New internal and external validation indices for clustering in Big Data

TESIS DOCTORAL

José María Luna Romera

Departamento de Lenguajes y Sistemas Informáticos
Escuela Técnica Superior de Ingeniería Informática
Universidad de Sevilla

Septiembre 2019

New internal and external validation indices for clustering in Big Data

Memoria que presenta para optar al título de Doctor en Informática con mención de Doctorado Internacional

José María Luna Romera

Dirigida por los Doctores

Dra. María del Mar Martínez Ballesteros
Dr. Jorge García Gutiérrez

Departamento de Lenguajes y Sistemas Informáticos
Escuela Técnica Superior de Ingeniería Informática
Universidad de Sevilla

Septiembre 2019

A mi familia
porque sin ella nunca hubiera llegado hasta aquí

Do. Or do not.
There is no try.
Yoda
Star Wars: The Empire Strikes Back

Tesis doctoral subvencionada por el Ministerio de Economía y Competitividad del Gobierno de España con los proyectos TIN2014-55894-C2-1-R y TIN2017-88209-C2-2-R, y por la Consejería de Innovación, Ciencia y Empresas de la Junta de Andalucía con el proyecto P11-TIC-7528.

Agradecimientos

La tesis doctoral aquí presentada es fruto del esfuerzo de varios años de dedicación en la cual aparezco como autor, pero sin duda, no hubiera sido posible sin el apoyo de todos los que me rodean y quisiera darles las gracias por tanto.

Quisiera agradecer en primer lugar a mis directores de tesis María y Jorge por todos sus consejos, y horas de dedicación y esfuerzo a lo largo de estos años para conseguir terminar esta tesis. Además, quisiera dar las gracias a Pepe por haberme acompañado durante esta etapa, aconsejándome y sabiéndome guiar por el mejor camino posible, muchas gracias.

Deseo además dar las gracias a mis compañeros del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla, y en especial a José Antonio Fábregas, por todo lo que ha supuesto desde que llegó, porque sin duda, marcó un punto de inflexión en este trabajo, ayudándome en todo lo que he "nezezitado", y parte de culpa de que haya llegado a escribir esta tesis es suya. Además, y no menos importante, quisiera dar las gracias a Álvaro, José David, Manuel, Pedro, Belén y Laura por estar acompañándome cada día y reinventar los desayunos.

Quisiera agradecer también a mis compañeros de la Universidad Pablo de Olavide, ya que de alguna manera allí empezó todo y tengo la suerte de poder contar con ellos cuando lo requiera. Quisiera además hacer una mención muy especial a Paco y a Alicia, porque desde que me gradué en aquella universidad han seguido contando conmigo y ellos han sido los partícipes principales para que a día de hoy yo pudiera estar aquí. Gracias, de corazón.

Me gustaría agradecer también a Manuel Roldán por acogerme de la forma en la que lo hizo cuando estuve de estancia en Arizona. Gracias por darme de la oportunidad de conocerle a él y a su familia, y por hacerme pasar 3 meses como si estuviera en mi propia casa. Nos vemos muy pronto.

Gracias a todos mis amigos porque me apoyaron desde el principio, y en especial a Richard, Edu, Selu y Murillo, porque nunca tengo que pedirles nada cuando lo necesito, y son los que siempre están ahí, tanto en los buenos como en los no tan buenos momentos, y aunque no se lo haya dicho nunca, son parte esencial de mi vida.

Quisiera además darle las gracias a Miriam porque ha sabido estar ahí

en el día a día, animándome, siempre a mi lado, apoyándome de manera incondicional, y pensando siempre cuál era el mejor camino que debía tomar. Gracias. Sin duda, has sido un pilar fundamental para que haya llegado a terminar esta tesis y quisiera agradecerte todo cuanto has hecho por mí.

Finalmente, quisiera dar las gracias a toda mi familia, y en especial a mis padres y a mi hermana, y aprovecho estas líneas para agradecerles todo el amor, cariño y apoyo que me han dado a lo largo de toda mi vida, porque siempre me han apoyado en todas las decisiones y han sabido darme todo cuanto he necesitado. De verdad, gracias.

Resumen

Esta tesis, presentada como un compendio de artículos de investigación, analiza el concepto de índices de validación de clustering y aporta nuevas medidas de bondad para conjuntos de datos que podrían considerarse Big Data debido a su volumen. Además, estas medidas han sido aplicadas en proyectos reales y se propone su aplicación futura para mejorar algoritmos de clustering.

El clustering es una de las técnicas de aprendizaje automático no supervisado más usada. Esta técnica nos permite agrupar datos en clusters de manera que, aquellos datos que pertenezcan al mismo cluster tienen características o atributos con valores similares, y a su vez esos datos son disimilares respecto a aquellos que pertenecen a los otros clusters. La similitud de los datos viene dada normalmente por la cercanía en el espacio, teniendo en cuenta una función de distancia. En la literatura existen los llamados índices de validación de clustering, los cuales podríamos definir como medidas para cuantificar la calidad de un resultado de clustering. Estos índices se dividen en dos tipos: índices de validación internos, que miden la calidad del clustering en base a los atributos con los que se han construido los clusters; e índices de validación externos, que son aquellos que cuantifican la calidad del clustering a partir de atributos que no han intervenido en la construcción de los clusters, y que normalmente son de tipo nominal o etiquetas.

En esta memoria se proponen dos índices de validación internos para clustering basados en otros índices existentes en la literatura, que nos permiten trabajar con grandes cantidades de datos, ofreciéndonos los resultados en un tiempo razonable. Los índices propuestos han sido testeados en datasets sintéticos y comparados con otros índices de la literatura. Las conclusiones de este trabajo indican que estos índices ofrecen resultados muy prometedores frente a sus competidores.

Por otro lado, se ha diseñado un nuevo índice de validación externo de clustering basado en el test estadístico chi cuadrado. Este índice permite medir la calidad del clustering basando el resultado en cómo han quedado distribuidos los clusters respecto a una etiqueta dada en la distribución. Los resultados de este índice muestran una mejora significativa frente a otros índices externos de la literatura y en datasets de diferentes dimensiones y

características.

Además, estos índices propuestos han sido aplicados en tres proyectos con datos reales cuyas publicaciones están incluidas en esta tesis doctoral. Para el primer proyecto se ha desarrollado una metodología para analizar el consumo eléctrico de los edificios de una smart city. Para ello, se ha realizado un análisis de clustering óptimo aplicando los índices internos mencionados anteriormente. En el segundo proyecto se ha trabajado tanto los índices internos como con los externos para realizar un análisis comparativo del mercado laboral español en dos periodos económicos distintos. Este análisis se realizó usando datos del Ministerio de Trabajo, Migraciones y Seguridad Social, y los resultados podrían tenerse en cuenta para ayudar a la toma de decisión en mejoras de políticas de empleo. En el tercer proyecto se ha trabajado con datos de los clientes de una compañía eléctrica para caracterizar los tipos de consumidores que existen. En este estudio se han analizado los patrones de consumo para que las compañías eléctricas puedan ofertar nuevas tarifas a los consumidores, y éstos puedan adaptarse a estas tarifas con el objetivo de optimizar la generación de energía eliminando los picos de consumo que existen la actualidad.

Palabras Clave

Minería de datos, clustering, índices de validación, Big Data

Abstract

This thesis, presented as a compendium of research articles, analyses the concept of clustering validation indices and provides new measures of goodness for datasets that could be considered Big Data. In addition, these measures have been applied in real projects and their future application is proposed for the improvement of clustering algorithms.

Clustering is one of the most popular unsupervised machine learning techniques. This technique allows us to group data into clusters so that the instances that belong to the same cluster have characteristics or attributes with similar values, and are dissimilar to those that belong to the other clusters. The similarity of the data is normally given by the proximity in space, which is measured using a distance function. In the literature, there are so-called clustering validation indices, which can be defined as measures for the quantification of the quality of a clustering result. These indices are divided into two types: internal validation indices, which measure the quality of clustering based on the attributes with which the clusters have been built; and external validation indices, which are those that quantify the quality of clustering from attributes that have not intervened in the construction of the clusters, and that are normally of nominal type or labels.

In this doctoral thesis, two internal validation indices are proposed for clustering based on other indices existing in the literature, which enable large amounts of data to be handled, and provide the results in a reasonable time. The proposed indices have been tested with synthetic datasets and compared with other indices in the literature. The conclusions of this work indicate that these indices offer very promising results in comparison with their competitors.

On the other hand, a new external clustering validation index based on the chi-squared statistical test has been designed. This index enables the quality of the clustering to be measured by basing the result on how the clusters have been distributed with respect to a given label in the distribution. The results of this index show a significant improvement compared to other external indices in the literature when used with datasets of different dimensions and characteristics.

In addition, these proposed indices have been applied in three projects

with real data whose corresponding publications are included in this doctoral thesis. For the first project, a methodology has been developed to analyse the electrical consumption of buildings in a smart city. For this study, an optimal clustering analysis has been carried out by applying the aforementioned internal indices. In the second project, both internal and external indices have been applied in order to perform a comparative analysis of the Spanish labour market in two different economic periods. This analysis was carried out using data from the Ministry of Labour, Migration, and Social Security, and the results could be taken into account to help decision-making for the improvement of employment policies. In the third project, data from the customers of an electric company has been employed to characterise the different types of existing consumers. In this study, consumption patterns have been analysed so that electricity companies can offer new rates to consumers. Conclusions show that consumers could adapt their usage to these rates and hence the generation of energy could be optimised by eliminating the consumption peaks that currently exist.

Keywords
Data mining, clustering, validation indexes, Big Data

Índice

Agradecimientos xi

Resumen xiii

Abstract xv

I Memoria 1

1. Introducción 3
1.1. Contexto 3
1.2. Problema 5
1.3. Solución 7
1.4. Estructura de la memoria de la tesis doctoral 7

2. Discusión conjunta de los resultados 9
2.1. BD-Silhouette y BD-Dunn 9
2.2. Chi Index 12
2.3. Aplicaciones 14

II Publicaciones 17

3. Trabajos de investigación seleccionados 19

4. An approach to validity indices for clustering techniques in Big Data 21

5. External clustering validity index based on chi-squared statistical test 39

6. Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities 59

7. Analysis of the evolution of the Spanish labour market through unsupervised learning 81

8. Big-Data Analysis for Demand Response in a Smart Electricity Market 101

III Conclusiones y Trabajos Futuros 119

9. Conclusiones y Trabajos Futuros 121
9.1. Conclusiones 121
9.2. Trabajo futuro 123
9.3. Conclusions 124
9.4. Future Work 125

IV Apéndices 127

A. Curriculum 129
A.1. Revistas indexadas JCR 129
A.2. Otras Revistas 130
A.3. Conferencias Nacionales 130
A.4. Proyectos I+D+i 131
A.5. Estancias 132

Bibliography 133

Índice de figuras

2.1. Representación de 3 *clusters* junto a las distancias *inter-cluster* $d(C_1, C_0)$ e *intra-cluster* $d(X_1, C_2)$, y el centroide global C_0 . 10
2.2. Representación de un conjunto de datos donde los círculos son puntos y los colores la clase a la que pertenecen. 14
2.3. Representación de la solución de *clustering* para $k = 2$ hasta 4. 14

Índice de Tablas

1.1. Resumen de artículos de investigación presentados. 8

2.1. Chi Index desde $k = 2$ hasta 4. 14

Parte I

Memoria

Capítulo 1

Introducción

Nunca perdáis contacto con el suelo; porque sólo así tendréis una idea aproximada de vuestra estatura

Antonio Machado
Juan de Mairena

1.1. Contexto

El *clustering* es una de las técnicas de aprendizaje automático no supervisado existentes dentro de la minería de datos. El objetivo del *clustering* es el de separar los datos de un conjunto en subconjuntos, llamados *clusters*, de manera que aquellos datos que pertenezcan a un mismo *cluster* sean similares, y a su vez disimilares a los de otros *clusters* [30]. De esta forma, el *clustering* crea subconjuntos de datos que comparten valores de atributos similares y que previamente eran desconocidos.

Dentro de la minería de datos, el *clustering* se define como una función de aprendizaje no supervisado, debido a que en el análisis de los datos no interviene ningún tipo de etiqueta o clase [24]. Podemos encontrar diferentes aplicaciones de *clustering* en la literatura, como por ejemplo la detección de outliers, ya que se podrían considerar aquellos puntos que queden más alejados a los *clusters* principales [41], el uso como herramienta para dividir un problema en pequeños subconjuntos y tratar individualmente a los *clusters* resultantes, incluso podríamos hacer uso de la etiqueta que proporciona el *clustering* para su posterior uso en la aplicación de técnicas de aprendizaje supervisadas. Recientemente se han realizado estudios aplicando técnicas de *clustering* en diversas áreas de conocimiento como en energía [44], química [69], medicina [13] o biología [11].

En la literatura podemos encontrar una gran cantidad de algoritmos, y aunque en algunos casos no es trivial hacer una categorización de ellos, los algoritmos podrían clasificarse de la siguiente forma [17]:

- ***Partitioning Methods***: dado un conjunto de n objetos, este tipo de métodos agrupa los objetos en k *clusters*, de manera que $k \leq n$ y cada *cluster* tiene al menos un objeto. Gran parte de los métodos de particionado forman los *clusters* basándose en distancias, de forma que, se asignan inicialmente k *clusters*, y se va iterando al cambiar los objetos de *clusters* hasta conseguir una solución donde cada objeto esté en su *cluster* más cercano [29, 51, 63]. Dentro de esta categoría podemos encontrar algoritmos como el k-means y el k-medoids.

- ***Hierarchical Methods***: este tipo de algoritmos realizan una descomposición jerárquica entre los objetos del conjunto. Estos métodos se podrían dividir a su vez en dos subtipos: aglomerativos, donde se va construyendo la jerarquía teniendo en cuenta todos los objetos de manera individual hasta acabar con un solo *cluster*; y divisivos, donde se parte de todos los objetos del conjunto en un mismo *cluster*, y desde este punto se van haciendo divisiones o descomposiciones de los mismos hasta que cada objeto quede en un *cluster* independiente [37, 60, 67].

- ***Density-Based Methods***: estos algoritmos construyen una solución de *clustering* en la que dados unos *clusters* iniciales, éstos van cambiando su forma en función de la densidad de los puntos que tienen alrededor en base a un umbral. Este tipo de método puede ser similar a los de particionado, solo que en estos casos las soluciones no tienen necesariamente forma esférica, sino que va adoptando una forma irregular a medida que van avanzando las iteraciones. A esta categoría pertenecen algoritmos como el DBSCAN [53] o el OPTICS [4].

- ***Grid-Based Methods***: estos métodos suelen aplicarse cuando el espacio del conjunto de datos es demasiado grande, ya que distribuyen los objetos en una cuadrícula, y el *clustering* se realiza directamente a cada celda dividiendo el problema de dimensionalidad [21, 62].

La evaluación de los resultados es una de las tareas más importantes y difíciles del *clustering*. Al aplicar diferentes algoritmos a un mismo conjunto de datos obtendremos diferentes soluciones de *clustering* y medir la calidad del *clustering* es tan importante como el método en sí [49]. Para evaluar qué solución de *clustering* es mejor, debemos tener en cuenta al menos dos factores: el número de *clusters*, ya que en general, el *clustering* depende directamente del número de *clusters* que definamos en su aplicación, y el resultado variará en función de este parámetro; y medir la calidad del *clustering* una vez obtenemos el resultado aplicando los llamados Índices de Validación de *clustering* (CVI, del inglés *Clustering Validation Indices*). Los CVI son unas medidas tomadas en base al resultado del *clustering* que cuantifican cómo han quedado los puntos distribuidos a través de los *clusters*. Existen numerosas medidas en la literatura, que se clasifican en dos tipos [24]:

- Índices Internos: medidas en las que se cuantifica la calidad del clustering en base a los atributos que se han usado para crear la solución. En general, este tipo de índice mide la distancia que existe entre los puntos que pertenecen a un mismo *cluster* (compacidad), y la separación que existe entre los distintos *clusters* [8, 15, 54]. Estos índices buscan una solución donde haya un alto grado de compacidad de los *clusters*, y a su vez los *clusters* estén lo más separados posible entre ellos. Este tipo de índice es el único que podemos aplicar a cualquier conjunto de datos, ya que hace uso de los propios atributos para construir la medida de calidad. En la literatura se pueden encontrar numerosos trabajos donde este tipo de índice ha sido usado [43, 23, 25, 28, 33, 59, 64], y entre los más utilizados se pueden encontrar: *Maximum Cluster Diameter* [26], *Average Within-Cluster Distance* [54], *Average Between-Cluster Distance* [54], *Silhouette* [50], *Dunn* [16], *Davies-Bouldin* [12] y *Calinski-Harabasz* [9].

- Índices Externos: estos índices miden la calidad del *clustering* en función de una etiqueta externa a la construcción del *clustering* [32, 34, 65, 68]. Estas medidas hacen una comparativa entre un ground truth dado y el *clustering* [2]. Este tipo de medidas no siempre se puede aplicar ya que dependerá de la disponibilidad del *ground truth*. En la literatura podemos encontrar diferentes índices de este tipo y existe una categorización hecha en función de cómo miden la calidad del *clustering*:

  - Entre los índices *set matching*, basados en la relación que existe entre dos soluciones de *clustering*, se encuentran *purity* [70], *F-measure* [31], *Criterion H* [40], *CSI* [19], *PSI* [47], y *Goodman-Kruskal* [22].
  - La categoría de *pair-counting* se basa en la comparación entre el número de objetos con la misma etiqueta dentro del mismo *clusters*. En esta categoría se encuentran: *Rand index* [46], *adjusted Rand index* [61], *Jaccard* [55], *Fowlkes-Mallows* [18], *Hubert Statistic* [27], y *Minkowski score* [7].
  - Además, se pueden encontrar índices basados en teoría de la información, como *entropy* [70], *variation of information* [39], y *mutual information* [6].

1.2. Problema

Hoy en día, cualquier dispositivo que nos rodea está generando datos constantemente. Se estima que para el 2020 el mundo digital pesará alrededor de 44 zettabytes, una cifra difícil de visualizar incluso si la pasamos a una medida con la que estemos más familiarizada ($4,4 * 10^{10}$ terabytes).

La gestión de esta cantidad de información ha supuesto un nuevo desafío para la comunidad científica debido a que algunas de las técnicas de análisis de datos que se usaban normalmente no están preparadas para trabajar con grandes cantidades de datos, y esta situación ha supuesto que algunas tengan que ser rediseñadas. Es por ello que, durante los últimos años, han ido surgiendo nuevas tecnologías especializadas en la gestión de esta gran cantidad de información [1, 38, 52].

Apache Hadoop [14] fue uno de los primeros *frameworks* que permitió el procesamiento de grandes cantidades de datos con un tiempo razonable de ejecución. Hadoop permite trabajar con *clusters* de ordenadores usando los modelos del paradigma de programación de *Google MapReduce* [14, 20]. No obstante, aunque *MapReduce* permitía gestionar grandes cantidades de datos, tenía su mayor inconveniente en los procesos de lectura y escritura, ya que realizaba estas operaciones directamente en disco y reducía considerablemente la velocidad de procesado. Como solución a este problema surgió *Apache Spark* [56], que resolvía las limitaciones de escritura-lectura de disco almacenando los datos en memoria, lo cual aceleraba el procesamiento de los datos, y lo hacía entre 10 y 100 veces más rápido que MapReduce [56]. Además, *Apache Spark* introdujo una nueva estructura de datos, llamada *resilient distributed dataset*(RDD), la cual fue especialmente diseñada para computación paralela porque almacena los resultados en memoria con el objetivo de procesar grandes cantidades de datos [66]. Por otra parte, *Apache Spark* incluye una librería de machine learning (MLlib) [57] con un conjunto de algoritmos para clasificación, regresión, árboles de decisión y *clustering*. En la literatura podemos encontrar nuevas aportaciones haciendo uso de esta librería [5, 42, 45, 58]

Algunas técnicas de *clustering*, como las categorizadas dentro de los métodos de particionado, necesitan como parámetro de entrada el número de *clusters* k en los que vamos a dividir el conjunto de datos. Como se ha comentado en la Sección 1.1, el resultado del *clustering* variará en función del valor de k, por tanto debemos optimizar este parámetro para conseguir la mejor solución de *clustering* posible. En la literatura se usan los CVI para mejorar los resultados de *clustering* basándonos en el k, sin embargo, estos éstos índices presentan ciertas limitaciones a la hora de trabajar con grandes cantidades de datos debido a su complejidad algorítmica [35].

Por otra parte, los CVI existentes en la literatura ofrecen un valor por cada k que estemos midiendo. Estos valores describen una curva, y cada índice ofrece el resultado óptimo teniendo en cuenta diferentes criterios como los mínimos o máximos locales, o aplicando el método del codo. Además, los resultados aportados por estos índices necesitan ser interpretados para obtener un buen resultado de *clustering*, y esto podría conllevar a error [3, 8, 10, 34, 48, 65, 68].

1.3. Solución

El objetivo de esta tesis doctoral es crear una solución de análisis de *clustering* para grandes cantidades de datos con el fin de obtener un resultado de *clustering* óptimo basándonos en CVI internos, y además obtener un CVI externo que ofrezca mejores resultados que los índices de la literatura y cuya solución no necesite ser interpretada. Por lo tanto, las soluciones aportadas en esta tesis doctoral podrían dividirse en dos partes:

- En primer lugar, se va a obtener una solución de *clustering* óptima teniendo en cuenta únicamente las características con los que se ha construido el *clustering*, es decir, considerando únicamente los atributos que se usan para agrupar los objetos. En este caso, se presentan dos CVI internos basados en las definiciones de los índices tradicionales, Silhouette [50] y Dunn [16], simplificando su implementación para poder tratar con grandes cantidades de datos (*Big Data*). Estos índices, denominados BD-Silhouette y BD-Dunn [35] muestran cuál es la solución de *clustering* óptima en un conjunto de datos que podría considerarse *Big Data*. Además, estos índices obtienen resultados prometedores en un tiempo de ejecución razonable sin reducir la pérdida en la precisión de los resultados.
- Por otra parte, se presenta un CVI externo, llamado Chi Index [36], basado en el test estadístico de chi cuadrado que ofrece una solución de *clustering* óptima sin necesidad de interpretar su resultado. De esta manera, Chi Index optimiza el resultado del *clustering*, de manera que, los *clusters* estarán compuestos por el menor número de distintas clases posible, y a su vez las clases estarán lo menos distribuidas a través de los *clusters*.

1.4. Estructura de la memoria de la tesis doctoral

Esta memoria de tesis doctoral por compendio de artículos está dividida en tres partes tal y como se detalla a continuación:

- En la Parte I se hace una introducción y se detallan los problemas y soluciones con los que se han trabajado en esta tesis. Asimismo, en esta parte se incluye la aplicación de las soluciones propuestas a datos de proyectos reales.
- En la Parte II se presentan los artículos de investigación derivados de esta tesis doctoral. Esta parte está dividida en capítulos, y cada uno corresponde a un artículo de investigación presentado.
- Finalmente, en la Parte III se exponen las conclusiones finales y trabajos futuros, así como el CV del candidato a doctor.

Un resumen de los artículos de investigación presentados en esta tesis doctoral se pueden encontrar en la Tabla 1.1:

Título	Revista	F.I. (JCR/SJR)	Ranking (JCR/SJR)
An approach to validity indices for clustering techniques in Big Data	Progress in Artificial Intelligence 2018	— / 0.513	- / Q2
External Clustering Validity Index based on chi-squared statistical test	Information Sciences 2019	5.524 / 1.62	Q1 / Q1
Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities	Energies 2018	2.707 / 0.67	Q3 / Q1
Analysis of the evolution of the Spanish labour market through unsupervised learning	IEEE Access (En revisión) 2019	4.098 / 0.61	Q1 / Q1
Big-Data Analysis for Demand Response in a Smart Electricity Market	IEEE Access (En revisión) 2019	4.098 / 0.61	Q1 / Q1

Tabla 1.1: Resumen de artículos de investigación presentados.

Capítulo 2

Discusión conjunta de los resultados

Estas obras no son mías porque las escriba yo sino porque yo he puesto sus fundamentos y razonamientos

Alfonso X el Sabio

En esta sección se presenta un resumen de las diferentes propuestas que se incluyen en los capítulos de la Parte II con el objetivo de dar una visión global de los artículos presentados en esta tesis doctoral.

2.1. BD-Silhouette y BD-Dunn

Como se ha comentado anteriormente en la Sección 1.2, los CVI tradicionales presentan ciertas limitaciones a la hora de trabajar con grandes cantidades de datos. En el artículo presentado en el Capítulo 4, se introducen dos nuevos CVI basados en índices de la literatura pero rediseñados para que sean capaces de trabajar con Big Data, ofreciendo el resultado de *clustering* óptimo en un tiempo de ejecución razonable.

En este artículo se escogieron los índices de Silhouette [50] y Dunn [16] entre 7 CVI de la literatura, ya que fueron los que mejor resultado obtuvieron en la experimentación llevada a cabo en [35]. Silhouette y Dunn están definidos en la literatura como sigue:

Sea Ω el espacio de los objetos con una distancia d. Entonces $\{A_k\}_{k=1..N}$ es un conjunto de *clusters* de manera que $\bigcup_k A_k = \Omega$, y $A_i \cap A_j = \emptyset \quad \forall i \neq j$. C_k es el centroide de A_k, y C_0 el centroide de Ω. Sea $x_i \in A_k$, la distancia de x_i a su propio *cluster* A_k viene definida por:

$$a_k(x_i) = \frac{1}{|A_k| - 1} \sum_{\substack{x_j \in A_k \\ j \neq i}} d(x_i, x_j) \tag{2.1}$$

Donde $a_k(x_i)$ representa la disimilitud de x_i al resto de puntos dentro del mismo *cluster* k, y

$$b_k(x_i) = \min_{j=i..N}\{a_j(x_i),\ j \neq k\} \tag{2.2}$$

$b_k(x_i)$ es el mínimo de la disimilitud media desde $x_i \in A_k$ a los puntos en los otros *clusters*.

- **Silhouette** se define en [50] como (Ec. 2.4):

$$s_k(x_i) = \frac{b_k(x_i) - a_k(x_i)}{max\{a_k(x_i), b_k(x_i)\}} \tag{2.3}$$

$$Silhouette = \frac{1}{|\Omega|}\sum_{x_i \in \Omega} s_k(x_i) \tag{2.4}$$

- **Dunn** se define en [16] como:

$$Dunn = \frac{\min\limits_{k=1..N}\{d(C_k, C_j),\ k \neq j\}}{\max\limits_{k=1..N}\{\max d(x_i, x_j), i \neq j,\ x_i, x_j \in A_k\}} \tag{2.5}$$

para un número de *clusters* N.

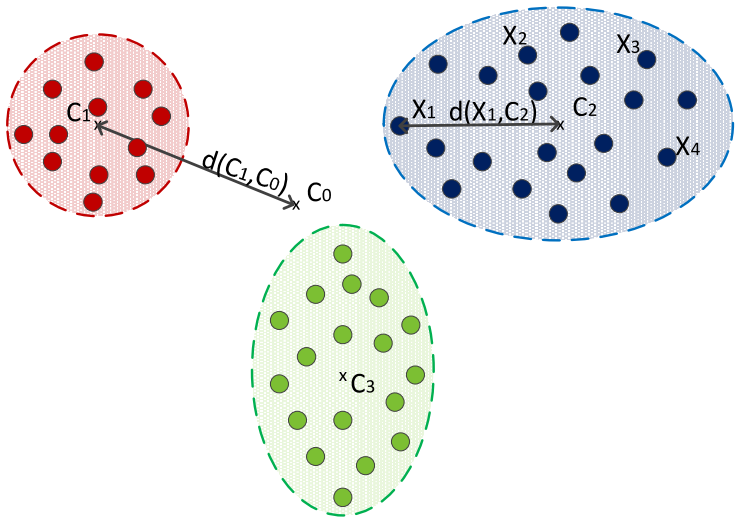


Figura 2.1: Representación de 3 *clusters* junto a las distancias *inter-cluster* $d(C_1, C_0)$ e *intra-cluster* $d(X_1, C_2)$, y el centroide global C_0

Las distancias descritas por estos índices están representadas gráficamente en la Figura 2.1. Como podemos observar, ambos índices tienen

como medida el cálculo de las distancias de cada punto a todos los puntos del *cluster* al que pertenece. Esto hace que estos algoritmos tengan complejidad cuadrática, lo cual supone un elevado coste computacional, así como la imposibilidad de paralelizar estos procesos haciendo uso de la tecnología Big Data.

Por estas razones surgen nuestra propuesta BD-Silhouette y BD-Dunn, los cuales reducen la complejidad de sus versiones tradicionales, sustituyendo los cálculos de distancias de punto a punto por los de las distancias a los centroides de los *clusters* a los que pertenece [35]. A continuación se detallan cada uno de estas medidas:

BD-Silhouette viene dado por las distancias *intra-cluster* e *inter-cluster*:

La distancia *inter-cluster* (Eq 2.6) es la media de las distancias entre cada centroide al centroide global C_0:

$$inter\text{-}cluster = \frac{1}{N}\sum_{k=1}^{N} d(C_k, C_0) \tag{2.6}$$

La distancia *intra-cluster* (Eq 2.8) es la media de las distancias entre los puntos y el centroide del *cluster* al que pertenece (Eq 2.7).

$$r_k = \frac{1}{|A_k|}\sum_{x_i \in A_k} d(x_i, C_k) \tag{2.7}$$

$$intra\text{-}cluster = \frac{1}{|N|}\sum_{x_i \in A_k}^{N} r_k \tag{2.8}$$

BD-Silhouette (Ec. 2.9) se define como el ratio entre la diferencia de las distancias *inter-cluster* e *intra-cluster*, y el máximo de ellos:

$$BD\text{-}Silhouette = \frac{inter\text{-}cluster - intra\text{-}cluster}{max\{inter\text{-}cluster, intra\text{-}cluster\}} \tag{2.9}$$

BD-Silhouette devuelve un valor en el rango $(-1, 1)$, y variará en función de la consistencia de los *clusters* y la separación que exista entre ellos. En los casos en los que el k sea mayor, hará que la distancia *intra-cluster* sea menor ya que los puntos del conjunto de datos tenderán a estar más compactos. Por lo tanto, BD-Silhouette valdrá -1 si el conjunto completo se agrupa en un único *cluster*, y tenderá a 1 cuando se vayan incrementando los *clusters*. En el caso extremo de que cada objeto sea un *cluster* BD-Silhouette valdrá 1. Por lo tanto, tomaremos como valor óptimo de BD-Silhouette el primer máximo local de

los k calculados, ya que este valor maximizará la coherencia del *cluster* con el menor k posible.

El rediseño para **BD-Dunn** se podría considerar similar. Este índice simplifica el índice original para facilitar la computación en Big Data ya que no calcula las distancias entre los pares de puntos, sino que se sustituye con la distancia a su centroide en el caso de las distancias entre puntos de un mismo *cluster*, y la distancia a un centroide global para el caso de las distancias *inter-cluster*. De manera que BD-Dunn (Eq 2.10) se define como el ratio entre el mínimo de las distancias de los centroides al centroide global, y el máximo de las distancias de cada punto al centroide del *cluster* al que pertenece:

$$BD\text{-}Dunn = \frac{\min\limits_{k=1..N}\{d(C_k, C_0)\}}{\max\limits_{k=1..N}\max\limits_{x_i \in A_k}\{d(x_i, C_k)\}} \tag{2.10}$$

En este caso, BD-Dunn valdrá 0 si todos los puntos son agrupados en un mismo *cluster*. Sin embargo, BD-Dunn tenderá a infinito mientras el número de *clusters* vaya incrementándose, ya que si llegáramos a agrupar cada punto en un *cluster* diferente no se podría calcular el valor de BD-Dunn porque el denominador de su función valdría cero.

2.2. Chi Index

Los índices externos miden la calidad de una solución de *clustering* basándose en cómo han quedado distribuidas las instancias en función de un atributo que no ha intervenido en la construcción del *clustering*. Estos índices, en general, realizan una comparativa entre el *clustering* y el *ground-truth*. Los índices de la literatura, generan un valor por cada k, y para obtener la solución de *clustering* óptima necesitan que su resultado sea interpretado a partir de los valores generados. Los índices de la literatura indican el *clustering* óptimo a partir de mínimos o máximos locales, o siguiendo el método del codo. Estos resultados pueden conllevar a conclusiones erróneas debido a que necesitan de una interpretación adicional. Con esta motivación se propone Chi Index, un índice de validación externo de *clustering* basado el test estadístico chi cuadrado de independencia de variables cualitativas y cuyo resultado mejora significativamente en tasa de acierto y error a 15 CVI de la literatura [36]. Chi Index se define formalmente como:

$$chi\,index(k) = row_{norm}(k) + col_{norm}(k) - |row_{norm}(k) - col_{norm}(k)| \tag{2.11}$$

donde

$$row_{norm}(k) = \frac{\chi^2_{row}(k)}{\chi^2_{row_{max}}} \tag{2.12}$$

$$col_{norm}(k) = \frac{\chi^2_{column}(k)}{\chi^2_{column_{max}}} \tag{2.13}$$

$$\chi^2_{row}(k) = \sum_{i}^{r}\sum_{j}^{c} \frac{(\frac{n_{ij}}{n_{i\cdot}} - \frac{N_{\cdot j}}{r})^2}{\frac{N_{\cdot j}}{r}} \tag{2.14}$$

$$\chi^2_{column}(k) = \sum_{i}^{r}\sum_{j}^{c} \frac{(\frac{n_{ij}}{n_{\cdot j}} - \frac{N_{i\cdot}}{c})^2}{\frac{N_{i\cdot}}{c}} \tag{2.15}$$

$$N_{i\cdot} = \sum_{j}^{c} \frac{n_{ij}}{n_{\cdot j}} \tag{2.16}$$

$$N_{\cdot j} = \sum_{i}^{r} \frac{n_{ij}}{n_{i\cdot}} \tag{2.17}$$

donde n_{ij} es el número de elementos del *cluster* i para la clase j, $n_{i\cdot}$ es el total de número de elementos en el *cluster* i, $n_{\cdot j}$ se corresponde con el total del número de elementos en la clase j, y n es el total de elementos en el *dataset* completo.

$$\chi^2_{row_{max}} = \begin{cases} 100 \cdot r \cdot (r-1) & r \leq c \\ 100 \cdot r \cdot (c-1) & r > c \end{cases} \tag{2.18}$$

$$\chi^2_{column_{max}} = \begin{cases} 100 \cdot c \cdot (r-1) & r \leq c \\ 100 \cdot c \cdot (c-1) & r > c \end{cases} \tag{2.19}$$

donde r y c son el número de filas y columnas respectivamente.

Chi Index toma un valor en el intervalo $[0, 2]$, donde 0 sería la peor solución de *clustering*, y 2 el mejor valor que Chi Index puede alcanzar. Por lo tanto, el valor óptimo de k viene dado por:

$$k^* =_k chi\, index(k) \tag{2.20}$$

Chi Index busca la mejor solución de *clustering* de manera que los *clusters* tengan la menor diversidad posible de clases, y a su vez las clases estén lo mejor distribuidas a través de los *clusters*. Tomemos la Figura 2.2 como ejemplo donde cada círculo es un punto en nuestro conjunto de datos, y los colores representan las clases a las que pertenece cada punto.

Si aplicamos *clustering* a este conjunto de datos desde $k = 2$ hasta $k = 4$, obtendríamos las soluciones representadas en la Figura 2.3. A simple vista

Figura 2.2: Representación de un conjunto de datos donde los círculos son puntos y los colores la clase a la que pertenecen.

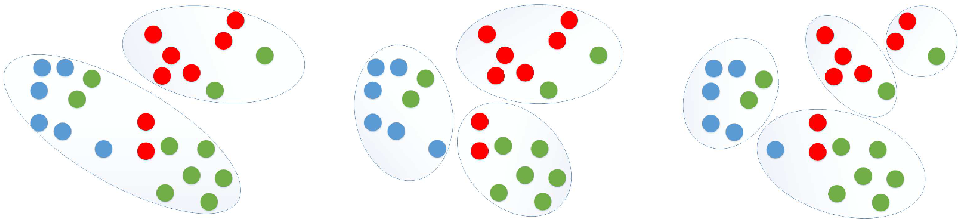


Figura 2.3: Representación de la solución de *clustering* para $k = 2$ hasta 4.

no es algo trivial ver cual de las tres soluciones de *clustering* es mejor, sin embargo, al aplicar Chi Index obtendremos un valor del índice para cada k.

La tabla 2.1 muestra los resultados de Chi Index en nuestro ejemplo. Aquí podemos observar que Chi Index alcanza su valor óptimo para $k = 3$, por tanto, podemos afirmar que es la mejor solución de *clustering*.

k	χ^2_{row}	χ^2_{column}	$\chi^2_{row_{max}}$	$\chi^2_{column_{max}}$	$chi\ index(k)$
2	89.01	139.40	200	300	0.890
3	277.50	299.38	600	600	**0.925**
4	304.05	237.21	800	600	0.760

Tabla 2.1: Chi Index desde $k = 2$ hasta 4.

2.3. Aplicaciones

Los índices comentados anteriormente en las Secciones 2.1 y 2.2 han sido aplicados en diferentes proyectos de investigación pudiendo trabajar con datos reales y solucionando problemas de diferente índole.

El primer problema que se abordó fue publicado en un artículo en colaboración con el grupo de investigación de la Universidad Pablo de Olavide (UPO) [44] para trabajar con datos del consumo eléctrico de dicha universidad. En este trabajo se propuso una metodología para analizar el consumo

consumo eléctrico de una smart city. Para ello se ha trabajado con datos reales del consumo eléctrico de los edificios de la Universidad Pablo de Olavide, que han sido preprocesados para aplicar técnicas de *clustering* con tecnología Big Data. Para este análisis, ante la imposibilidad de aplicar los tradicionales CVI debido a la cantidad de datos con los que se trabajaba, se aplicaron los índices BD-Silhouette y BD-Dunn [35] para descubrir el número óptimo de *clusters* del conjunto de datos. Una vez calculado el número óptimo de *clusters*, se hizo un análisis de los *clusters* caracterizando los resultados. En este caso se calcularon las tablas de contingencia de los *clusters* tomando como etiquetas los edificios, las estaciones del año, los día de la semana y si era laborable o no. De esta manera obtuvimos una completa caracterización de los *clusters*, obteniendo información relevante del consumo eléctrico de los diferentes edificios de la UPO. Esta metodología podría ser aplicada a datos de consumo eléctrico de una smart city con vistas a que los resultados puedan ser tratados para la ayuda a la toma de decisiones de un gobierno o una administración pública.

El segundo de los problemas reales surgió de una colaboración con compañeros del Departamento de Organización Industrial de la Universidad de Sevilla y del Departamento de Economía de la UPO. En este proyecto se trabajó con datos del Ministerio de Trabajo, Migraciones y Seguridad Social, en concreto con la Muestra Continua de Vidas Laborales. El objetivo de este proyecto fue el de descubrir cómo se organiza el mercado laboral teniendo en cuenta datos de las colocaciones de los trabajadores en dos periodos económicos bien diferenciados: 2011-2013, años correspondientes a la crisis económica, y 2014-2016, periodo que comprende los primeros años de recuperación. En este análisis se trataron 1,9 y 2,4 millones de colocaciones respectivamente, lo que podría considerarse un problema real de Big Data. En este análisis se aplicaron los dos tipos de índices descritos en las Secciones 2.1 y 2.2. Debido a la naturaleza y forma de los datos, los índices BD-Silhouette y BD-Dunn [35] no ofrecieron unos resultados lo suficientemente claros como para tenerlos en cuenta a la hora de analizar el *clustering*. Sin embargo, Chi Index [36] mostró unos claros resultados tomando la comunidad autónoma, la provincia, la actividad y la ocupación como clases. Finalmente, se analizó la solución de *clustering* dada por el k óptimo de la comunidad autónoma ya que por definición incluía la información de la provincia, su valor era superior al de las otras clases, y además era un número de *clusters* fácil de manejar e interpretar. El análisis de *clustering* se realizó tomando como número óptimo de *clusters* el ofrecido por el Chi Index. Los resultados del análisis fueron prometedores ya que es muestra la transformación del mercado laboral a través de los *clusters* creados en ambos periodos. Estos resultados podrían llegar a apoyar las decisiones económicas y políticas de las administraciones públicas con el objetivo de mejorar en calidad de política de empleo.

El tercer problema que se afrontó aplicando los índices de validación fue en colaboración con una compañía eléctrica europea, en un estudio sobre los hábitos de consumo de sus clientes y en cómo optimizar la energía producida en función de esos hábitos, con el objetivo de conseguir un ahorro económico para los clientes y a su vez un ahorro energético para las eléctricas que pueden conseguir aplanar la curva de la demanda. En este trabajo se utilizaron 1,8 TB de datos tomados de los contadores inteligentes de los clientes y se analizaron mediante técnicas de *clustering* los consumos anuales de los clientes. Para el cálculo del número óptimo de *clusters* se utilizaron los índices de BD-Silhouette y BD-Dunn, y concluyeron que existían 6 tipos de clientes diferentes. En los resultados obtenidos se puede observar que hay una clara distinción entre los hábitos de consumos de estos *clusters*, y además que estos consumos no están directamente relacionados con la potencia contratada, existiendo *clusters* de bajo consumo con una potencia contratada muy superior. Los resultados de este estudio nos permiten conocer el comportamiento de los consumidores para que las compañías eléctricas puedan acomodar sus tarifas a estos comportamientos y mejorar su respuesta a la demanda. Además, los consumidores podrían adaptarse a las nuevas tarifas con el fin de que no haya un exceso de generación de energía por parte de las compañías, ya que esto supondría un ahorro para ambas partes al eliminar los picos de consumo que las eléctricas sufren.

Parte II

Publicaciones

Capítulo 3

Trabajos de investigación seleccionados

La razón es el único don del cielo que compensa plenamente los males de la existencia humana

José María Blanco White
Cartas de España, III

En esta parte de la memoria se detallan los artículos de investigación incluidos en esta tesis doctoral dividido por capítulos:

- Capítulo 4: **An approach to validity indices for clustering techniques in Big Data**. José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros, José C. Riquelme. **Progress in Artificial Intelligence** , Volumen: 7(2), 91-94, 2018. DOI: `https://doi.org/10.1016/10.1007/s13748-017-0135-3`. F.I.(SJR-2018): 0.513.

- Capítulo 5: **External Clustering Validity Index based on chi-squared statistical test** José María Luna-Romera, María Martínez-Ballesteros, Jorge García-Gutiérrez, José C. Riquelme **Information Sciences**, Volumen: 487, 1-17, 2019. DOI: `https://doi.org/10.1016/j.ins.2019.02.046`. F.I.(JCR-2018): 5.524.

- Capítulo 6: **Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities**. Rubén Pérez-Chacón, José María Luna-Romera, Alicia Troncoso, Francisco Martínez-Álvarez, José C. Riquelme. **Energies**, Volumen: 11(3), 683, 2018. DOI: `https://doi.org/10.3390/en11030683`. F.I.(JCR-2018): 2.707.

- Capítulo 7: **Analysis of the evolution of the Spanish labour market through unsupervised learning** José María Luna-Romera,

Fernando Nuñez-Hernández, María Martínez-Ballesteros, José C. Riquelme, Carlos Usabiaga. **IEEE Access**. En revisión. F.I.(JCR-2018): 4.098.

- Capítulo 8: **Big-Data Analysis for Demand Response in a Smart Electricity Market** José Antonio Fábregas, José María Luna-Romera, David Gutiérrez-Avilés, José C. Riquelme. **IEEE Access**. En revisión. F.I.(JCR-2018): 4.098.

Capítulo 4

An approach to validity indices for clustering techniques in Big Data

Resumen

El clustering es una de las técnicas de aprendizaje no supervisado más usadas en minería de datos. Tiene como objetivo el de agrupar los datos por similitud, de manera que aquellos datos que pertenezcan al mismo grupo, o *cluster*, sean muy parecidos entre ellos, y a su vez los grupos presentan un grado de disimilitud. Para medir la bondad de un clustering, existen los llamados índices de validación de clustering, que nos permiten entender cómo han quedado distribuidos los datos por los *clusters*. En general, un buen resultado de clustering es aquel en el que los puntos que pertenecen a un mismo *cluster* son muy similares, y a la vez los puntos en diferentes *clusters* son distintos entre ellos. En la literatura existen numerosos CVI pero éstos índices presentan ciertas limitaciones a la hora de trabajar con grandes cantidades de datos debido a su complejidad computacional. En este artículo se presentan dos novedosos CVI para Big Data (BD-CVI) basados en los índices de la literatura Silhouette y Dunn, los cuales han sido diseñados para reducir la complejidad que éstos presentan y poder ofrecer el resultado en un tiempo de ejecución razonable. Para probar estos novedosos BD-CVIs, la experimentación se ha llevado a cabo usando 28 datasets sintéticos, de los cuales se conocía a priori el número óptimo de *clusters*. El tamaño de estos datasets varía en número de instancias y número de *clusters*, y van desde los 5 hasta los 11 *clusters*, y desde 5.000 hasta 11 millones de instancias. Para probar la efectividad de estos CVI se han aplicado los tests estadísticos de Friedman y un análisis post-hoc usando el procedimiento de Holm. Los resultados indican que BD-Silhouette y BD-Dunn son significativamente diferentes a Davies-Bouldin. Además, se realizó un estudio para medir los

tiempos de ejecución de estos BD-CVI y como se muestra en los resultados, nos permiten trabajar con grandes cantidades de datos ofreciéndonos resultados en un tiempo razonable. El artículo incluye un apéndice en el cual se detalla cómo se han contruído los datasets sintéticos para la experimentación. Además, queda disponible el código en Github para Spark tanto del generador de los datasets como de los BD-CVIs presentados en este artículo.

- Estado: Publicado en Progress in Artificial Intelligence (Springer) (2018), Volumen: 7(2), 91-94
- Índice de Impacto (SJR 2018): 0.513
- Área de Conocimiento:
  - Computer Science, Artificial Intelligence. Ranking - Q2
- Citas:
  - Scopus: 4
  - Google Scholar: 9
  - Web of Science: 1

Prog Artif Intell (2018) 7:81–94
https://doi.org/10.1007/s13748-017-0135-3

REGULAR PAPER

An approach to validity indices for clustering techniques in Big Data

José María Luna-Romera[1] · Jorge García-Gutiérrez[1] · María Martínez-Ballesteros[1] · José C. Riquelme Santos[1]

Received: 8 August 2017 / Accepted: 21 September 2017 / Published online: 5 October 2017

Abstract Clustering analysis is one of the most used Machine Learning techniques to discover groups among data objects. Some clustering methods require the number of clusters into which the data is going to be partitioned. There exist several cluster validity indices that help us to approximate the optimal number of clusters of the dataset. However, such indices are not suitable to deal with Big Data due to its size limitation and runtime costs. This paper presents two clustering validity indices that handle large amount of data in low computational time. Our indices are based on redefinitions of traditional indices by simplifying the intra-cluster distance calculation. Two types of tests have been carried out over 28 synthetic datasets to analyze the performance of the proposed indices. First, we test the indices with small and medium size datasets to verify that our indices have a similar effectiveness to the traditional ones. Subsequently, tests on datasets of up to 11 million records and 20 features have been executed to check their efficiency. The results show that both indices can handle Big Data in a very low computational time with an effectiveness similar to the traditional indices using Apache Spark framework.

Keywords Clustering · Big Data · Clustering validity indices · Intra-cluster distance

José María Luna-Romera
jmluna@us.es

Jorge García-Gutiérrez
jorgarcia@us.es

María Martínez-Ballesteros
mariamartinez@us.es

José C. Riquelme Santos
riquelme@us.es

[1] Department of Computer Languages and Systems, ETSII, University of Seville, Seville, Spain

1 Introduction

In the last few years, available data has been increased considerably. Medicine, electricity, business or biology are some areas where data has been quickly generated [4,7,13,29,31,40]. This information needs to be processed in order to discover knowledge, but traditional techniques are limited by the size of the data. This fact supposes a challenge to the research community because traditional machine learning methods cannot deal with large volume of data. Therefore, such learning techniques need to be redesigned to be able to handle Big Data.

Among the traditional techniques to discover knowledge, clustering can be useful to analyze large datasets with the aim at finding groups with similar behavior. Clustering is formally defined in [15] as the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures. Each clustering method generates different solutions on the same dataset. Clustering analysis is also applied to detect unknown associations within the data.

In particular, clustering techniques based on partitioning methods find the most suitable partition of the objects of the dataset into a given number of groups optimizing a chosen partitioning criterion. Nevertheless, such methods require the optimal number of clusters that the dataset is going to be partitioned. For this task, there exist cluster validity indices (CVI) that help to calculate it. The application and usability of these indices has been proven in several works in the literature [1–3,25]. However, the traditional indices are not suitable to deal with large datasets due to the high computational time costs and their inability to be parallelized. Traditional CVIs

use pairwise distances, so such CVIs will have quadratic complexity. The use of this kind of CVIs on large data could take much longer to compute the evaluation measure than running the clustering algorithm.

Nowadays, some frameworks are able to deal with Big Data. One of the first frameworks that allowed processing large datasets was Apache Hadoop [9]. Hadoop allows to work across clusters of computers using simple programming models based on Google's MapReduce paradigm [9,14]. Additionally, one of the most used Big Data projects is the open-source cluster computing framework named Apache Spark [34]. Spark appeared as alternative to solve memory limitation that MapReduce suffered. MapReduce reads and writes from hard drive, as a result, it slows down the processing speed. Spark reduces the number of read/write cycles to disk and stores intermediate data in faster logical RAM memory. It uses an structured data, named RDD, especially designed for parallel computing that caches results in memory for processing large amounts of data [38]. Apache Spark contains an scalable Machine Learning library (MLlib) with a set of algorithms to handle classification, regression, decision tree, recommendation systems and clustering techniques [35].

The purpose of this paper is to show the limitations of traditional clustering indices and to present novel validity indices that can tackle Big Data, henceforth named BD-CVI. In particular, the proposed indices are implemented using Apache Spark framework. A data generator application is also presented to ensure the composition of data and to test the performance of the proposed indices. K-means method was selected for testing the performance of these CVIs and BD-CVIs.

The remainder of this paper is organized as follows. Section 2 presents an outline of the background about clustering including a description of traditional indices and the state-of-the-art of Big Data clustering. Section 3 defines the BD-CVI proposed in this paper. Section 4 shows the experimental analysis and the obtained results. Section 5 reports the conclusions drawn by this paper. Lastly, "Appendix A" details the dataset generator application that was implemented to be used in the experimentation.

2 Related work

In this section, clustering analysis is formally defined and a general classification of the main clustering methods is also presented (Sect. 2.1). In addition, main CVIs are described and classified by categories (Sect. 2.2). Furthermore, we provide a brief overview of previous works related to clustering analysis in Big Data (Sect. 2.3).

2.1 Clustering methods

Cluster analysis can be used as a mechanism to achieve a custom vision of the distribution of the data, to observe the features of each cluster and to target on a particular subset of data for other analysis. It is also used as a preprocessing step for further algorithms, such as classification or features selection, which would deal with the detected clusters and the selected features. In some cases, clustering analysis is also called automatic classification since clustering is a collection of similar data objects on each cluster, so data objects within the same cluster can be managed as an implicit class. Clustering can be found in the literature as data segmentation in some applications because large datasets are partitioned into groups by their similarity. Another use of clustering is the outlier detection, that could be defined as that data object that is far away from any cluster and it may be more interesting to not to include it in any of them. Some applications of clustering for outlier detection can be found in [26].

Clustering analysis is considered a branch of statistics, and it has been widely studied as distance-based cluster analysis. Clustering analysis implementations based on K-means or K-medoids have been developed into many statistical analysis software packages and systems [18]. In Machine Learning, clustering is a method of unsupervised learning, so the data object has no class label information. Clustering learns by observation instead of learning by examples. Some of the active research topics are focused on the scalability of clustering methods [6], the effectiveness of types of data [32], high-dimensional clustering techniques [18], and methods for clustering mixing numerical and nominal data in large datasets.

There exist many clustering algorithms in the literature, so it is difficult to set a categorization. In many cases, an algorithm can be classified into several categories due to its features. However, most of the clustering techniques could be classified into the following categories [12]:

Partitioning methods Given a set of n objects, a partitioning method constructs k groups of data, where each partition represents a cluster and $k \leq n$. It splits the data objects into k clusters such each cluster must contains at least one data object. Most of these methods are distance-based, so given k, which is the number of groups to build, a partition method sets an initial solution. Then, it iterates and try to improve the solution by moving objects between the groups. Despite each clustering method takes its own criterion, a satisfying partitioning is where data objects in the same cluster are close and objects in different clusters are far away. Clustering methods usually work properly with spherical-shaped cluster when this operation is used [18].

Hierarchical methods This kind of methods create a hierarchical decomposition of the given set of data objects. It

successively groups the objects close to one another until all the data objects are merged into one. This kind of clustering methods leads to smaller computation costs by not having to worry about a combinatorial number of different choices due to once a step is done it can never be undone.

Density-based methods This methods iteratively build a cluster as long as the density, defined as the number of objects in a cluster, exceeds some threshold. This kind of methods is used to detect outliers and discovers random-shape clusters.

Grid-based methods This kind of methods compute the object space into a finite number of cells that form a grid structure. They are independent of size of the dataset and dependent on the number of cells in each dimension in the quantized space.

2.2 Cluster validity indices

As stated in the introduction, this paper is focused on partitioning clustering methods because they need the number of clusters into which the dataset is going to be partitioned. Knowing a priori a proper approximation to the number of cluster could be very useful for any clustering algorithm, especially hierarchical ones. For this task, several CVIs have been proposed in the literature. A summary of the most representative CVIs of each category are presented as follows [10,33]:

- *Indices measuring compactness of clusters* These indices measure both the distance between the points that belong to the same cluster and the compactness of them:

  *Maximum Cluster Diameter (Δ) [16] is the highest diameter among all the clusters in the dataset. It is calculated by the maximum distance between two points that belong to the same cluster.
  *Average Within-Cluster Distance (W) [33] measures the average distance of the points that belong to the same cluster.

- *Indices measuring separation between clusters* This category evaluates the separation of the clouds of points and grades it when there is a gap between them:

  *Average Between-Cluster Distance (β) [33] is the average distance of the points that are in different clusters.

- *Indices measuring relationships between compactness and separation* This category measures the ratio between the compactness of the clusters and the existing separation between them:

  *Silhouette [30] is a measure that sets how compacted are the points that belong to the same cluster against the separation between the clusters.

  *Dunn [11] measures the relation between the minimum inter-cluster distance and the maximum intra-cluster distance.

  *Davies and Bouldin [8] is a measure that uses data object quantities and features inherent to the dataset to set the compactness and separation of the clusters.

  *Calinski and Harabasz [5] is based on getting a relation between the inter-cluster distance and the intra-cluster distance.

2.3 Clustering in Big Data

Clustering analysis in Big Data has been the main focus of a lot of researchers in the last years. Some of the most relevant papers in this field are analyzed below.

Two C-means algorithms based on the canonical polyadic decomposition and the tensor-train network for clustering Big Data are proposed in [39]. They stated that the algorithms are suitable for Big Data clustering in Internet of Things systems with low-end devices since they can achieve a high compression rate for heterogeneous samples to save the memory space significantly.

Mohammed et al. [27] proposed a new cluster algorithm named FireflyClust, that it can deal with text documents in a hierarchical line. FireflyClust can handle Big Data, overcoming other methods such as Bisect K-means, hybrid Bisect K-means and Practical General Stochastic Clustering Method. In this case, the algorithm does not need the number of cluster as input parameter.

An effective K-means algorithm design was proposed in [37]. The algorithm is based on MapReduce programming model that acquires a fast detection speed with a high scale-up. However, no method was applied to identify the optimum number of clusters, so the algorithm was tested using different k number of clusters.

Jerome and ätönen [20] proposed a hierarchical clustering technique for classifying anomalies into clusters and providing information regarding the behavior of the anomaly cluster by analyzing its centroid in Big Data. Overall, it is easier to detect anomalies than finding out reasons for anomalous behavior. This technique was also used to determine the severity of the anomaly by using a failure significance metric.

A novel cluster center fast determination clustering algorithm for Big Data was proposed in [21]. The algorithm is based on the density and distance distribution of the data objects to determine the cluster center quickly by constructing the normal distribution function.

Kim et al. [22] suggested an optimized combinatorial clustering algorithm for noisy performance with random sampling for Big Data. The algorithm outperforms conventional approaches through various numerical and qualitative

thresholds such as mean and standard deviation of accuracy and computation speed.

Tong et al. [36] proposed Scalable Clustering Using Boundary Information, a highly flexible and scalable clustering scheme. To achieve this, such algorithm firstly identifies the border points of the dataset, and then it groups boundary points into suitable clusters and includes the rest points to their nearest border point. The obtained reports confirm similar results than the standard DBSCAN method, but such method is able to handle Big Data.

A novel method for assessing the robustness of clusters for partitioning algorithms is introduced in [23]. However, such method is not applied to Big Data, and moreover, the experiments have been carried out with supervised data where classes have been used as clusters.

3 Big Data indices

As stated in Introduction, the main purpose of this paper is to provide efficient and suitable CVIs able to deal with Big Data. The proposed BD-CVIs are approximations of traditional indices because those indices require high computational cost and they are unable to be parallelized. Firstly, traditional CVIs are described in Sect. 3.1. Secondly, the definition of BD-CVIs is in Sect. 3.2.

3.1 Traditional CVIs

From the traditional indices, we have selected those that had the best performance in the experiments (Sect. 4.3.3), and thus Silhouette and Dunn are detailed below:

Let Ω be the space of the objects with a given distance d.

Then, $\{A_k\}_{k=1...N}$ is a set of clusters so that $\bigcup_k A_k = \Omega$, and $A_i \cap A_j = \emptyset \quad \forall i \neq j$.

C_k is the centroid of A_k, and C_0 the centroid of Ω.

Let $x_i \in A_k$, the distance from x_i to the own cluster A_k is defined:

$$a_k(x_i) = \frac{1}{|A_k| - 1} \sum_{\substack{x_j \in A_k \\ j \neq i}} d(x_i, x_j) \tag{1}$$

where $a_k(x_i)$ represents the dissimilarity of x_i to all other points within the same cluster k and

$$b_k(x_i) = \min_{j=i...N} \{a_j(x_i),\ j \neq k\} \tag{2}$$

$b_k(x_i)$ is the smallest average dissimilarity of $x_i \in A_k$ to the points in other clusters.

– *Silhouette* is defined in [30] as (Eq. 4):

$$s_k(x_i) = \frac{b_k(x_i) - a_k(x_i)}{max\{a_k(x_i), b_k(x_i)\}} \tag{3}$$

$$Silhouette = \frac{1}{|\Omega|} \sum_{x_i \in \Omega} s_k(x_i) \tag{4}$$

Silhouette index ranges $[-1, 1]$, where good values are near 1 and -1 closer values are bad clustering solutions.

– *Dunn* index is defined in [11] and its purpose is to identify compact and well-separated clusters. For a given number of clusters N, the following equation defines the Dunn index:

$$Dunn = \frac{\min\limits_{k=1...N} \{d(C_k, C_j),\ k \neq j\}}{\max\limits_{k=1...N} \{\max d(x_i, x_j), i \neq j, x_i, x_j \in A_k\}} \tag{5}$$

In a compact and well-separated clusters dataset, the distances between the clusters are wide and the distances between the points of the same cluster are small. Hence, a high value of the Dunn index means a compact and well-separated clusters solution.

3.2 BD-CVIs

In this subsection, BD-Silhouette and BD-Dunn are going to be formally introduced:

– *BD-Silhouette* is defined by two approaches to intra-cluster and inter-cluster mean distances.
inter-cluster (Eq. 6) is the average of distances between each cluster centroid and global centroid C_0:

$$inter\text{-}cluster = \frac{1}{N} \sum_{k=1}^{N} d(C_k, C_0) \tag{6}$$

where C_0 is the center of the centroids of the clusters.
intra-cluster (Eq. 8) distance is defined as the average of the distances between each point to the centroid of the cluster to which it belongs (Eq. 7).

$$r_k = \frac{1}{|A_k|} \sum_{x_i \in A_k} d(x_i, C_k) \tag{7}$$

$$intra\text{-}cluster = \frac{1}{|N|} \sum_{x_i \in A_k}^{N} r_k \tag{8}$$

Traditional Silhouette index takes *intra-cluster* distance instead, that is defined by the average distance between

the points that belong to the same cluster (Eq. 1). The *intra-luster* distance is the main difference in BD-Silhouette.

Equation 9 represents BD-Silhouette that has been defined as the ratio between the difference of the *inter-cluster* and the *intra-cluster*, and the maximum of them.

$$BD\text{-}Silhouette = \frac{inter\text{-}cluster - intra\text{-}cluster}{max\{inter\text{-}cluster, intra\text{-}cluster\}} \tag{9}$$

BD-Silhouette returns a value in $(-1, 1)$, depending on the consistence of the cluster and the separation between them. The higher the cluster number is, the lower *intra-cluster* is because the points of the dataset tend to be more compact. BD-Silhouette takes the value -1 if a single cluster is defined for all the examples and tends to 1 when the number of clusters is increased. BD-Silhouette would be 1 in the extreme case of each data object being a cluster. Therefore, an optimal value for the number of clusters would be the first maximum of BD-Silhouette, which maximizes the coherence of the cluster with the lowest k possible.

– *BD-Dunn* simplifies the original Dunn index to facilitate its computation in Big Data, since it does not have to calculate in the denominator the distance between each pair of points of the dataset. On the contrary, original Dunn seeks the minimum distance between the centroids and the maximum distance between all the points that belong to the same cluster. Thus, BD-Dunn (Eq. 5) is the ratio between the minimum of the distances from the centroids to the global center and the maximum of the distances from each point in the set to its centroid.

$$BD{-}Dunn = \frac{\min\limits_{k=1\ldots N}\{d(C_k, C_0)\}}{\max\limits_{k=1\ldots N}\max\limits_{x_i \in A_k}\{d(x_i, C_k)\}} \tag{10}$$

BD-Dunn takes the value 0 if we define a single cluster for all the examples. However, BD-Dunn tends to infinity when the number of clusters increases. In the extreme case of each example belong to a different cluster, its value cannot be calculated because the denominator is zero.

Figure 1 illustrates a distribution of a dataset with 2 features and 3 clusters in different colors. The clusters are represented by circles, and the points are the red dots. Each cluster i has its centroid denoted by C_i, and the global centroid as C_0 is also represented. Blue cluster has also highlighted some points in the cluster. In the figure, *inter-cluster* distance is represented by the red cluster as $d(C_1, C_0)$ that measures the distance between the centroid of the red cluster and the global centroid. The *intra-cluster* distance is represented in the blue cluster as the distance between the point X_1 and its centroid C_2.

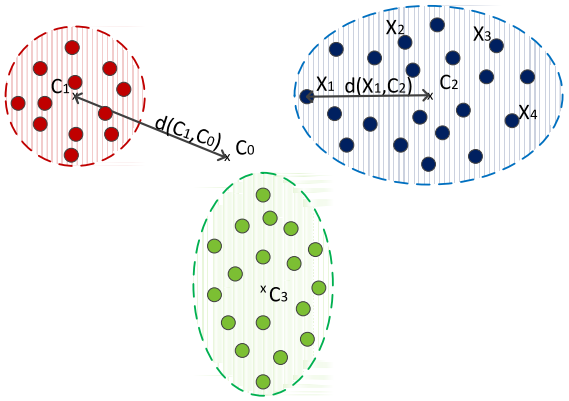


Fig. 1 Representation of 3 clusters with *inter-cluster* and *intra-cluster* distance and the global centroid (C_0)

As it happens with traditional CVIs, BD-CVIs return a value on each clustering solution. To get the optimum number of clusters of a large dataset, BD-CVI could be calculated on each clustering execution. The optimal number of clusters is chosen following a different criterion on each BD-CVI. BD-Silhouette and BD-Dunn are growing indices. BD-Silhouette reaches 1 when $k = N$, and BD-Dunn tends to infinity. Thus, in both BD-CVIs, the first maximum is a satisfactory solution because it maximizes the clustering coherence with the lowest number of clusters possible.

Figure 2 illustrates the graphical representation of the results of BD-Silhouette and BD-Dunn for a dataset with 5 clusters and 500,000 instances each. BD-Silhouette value increases with the number of clusters. Such index marks a possible optimal number of clusters when there is a change of trend in the values. In Fig. 2, BD-Silhouette is increased by the number of clusters until $k = 5$. This change of trend indicates that $k = 5$ may be an optimal number of clustering. BD-Dunn reveals the optimal number of clusters with the first maximum value of its plot. Figure 2 shows a first maximum value on $k = 5$, where the line of BD-Dunn increases with the number of clusters and decreases in $k = 6$. This inflection point indicates that $k = 5$ could be the optimal number of clusters.

4 Experimental study

The experimental setup and the results are detailed in this section. A comparative framework is also presented to test

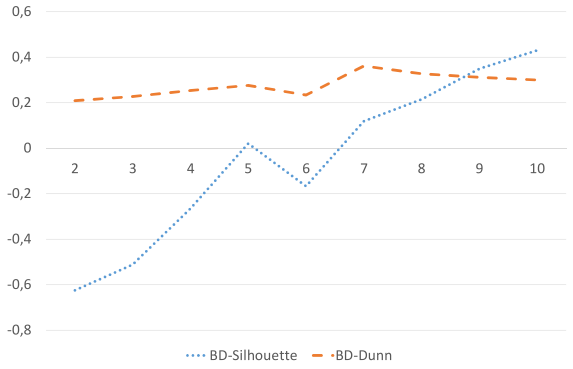


Fig. 2 BD-Silhouette and BD-Dunn results for a dataset with 5 clusters and 500,000 instances each

both CVIs and to test which CVIs and BD-CVIs have the best performance (Sects. 4.3.3, 4.4.3).

4.1 Experimental setup

4.1.1 Software and hardware

In this paper, we compared the results of traditional CVIs and the proposed BD-CVIs with the datasets described in Sect. 4.1.2. A clustering algorithm is required to test the performance of the proposed BD-CVIs. As stated in Sect. 2, K-means is a partitioning method that previously needs the number of clusters into which the dataset is going to be partitioned, so that this algorithm has been selected in the experimental study. In addition, it is the paradigmatic clustering algorithm [19] and it is one of the available algorithms in Spark MLlib [35]. In the case of traditional CVIs, we have used the K-means package available in the Weka Software developed in Java [17].

Two different execution environments were used in our experiments. On the first hand, traditional CVIs have been tested in the EC2 instances from Amazon Web Services (AWS) that count with Intel Xeon E5-2666 v3 (Haswell) processors, 3.75 GB RAM memory and enough hard disk to manage datasets originally stored in AWS S3. On the other hand, BD-CVIs were executed in AWS Elastic Map Reduce. 5 instances of $m3.xlarge$ that each one counts with Intel Xeon E5-2670 v2 (Ivy Bridge) processors with 4 vCPU, 15 GB RAM memory and 2 SSD of 40 GB were used.

4.1.2 Generated datasets

A total of 28 datasets have been used which are generated using the dataset generator application described in "Appendix A". In order to test the limits of the CVIs and the novel BD-CVIs, several combinations of number of clusters and number of instances were applied. Table 1 shows the main features of the generated datasets in the experiments. There are 4 different groups of datasets with 5, 7, 9 and 11 clusters. Each group of datasets contains 7 datasets, with 1000, 2000, 5000, 10,000, 100,000, 500,000 and 1 million instances per cluster. Thus, a dataset with 5 clusters and 1000 instances per cluster has a total of 5000 instances. The datasets are easily identified following the next pattern: $C - N\{k, M\}$ where C is the optimal number of clusters of the dataset, N is the number of instances of each cluster multiplied by a thousand if it is followed by a k, or by a million if it is followed by a M. For example, 5–10k dataset has 5 clusters and each one contains 10,000 instances, so it contains a total of 50,000 instances. All the datasets were created with 20 features, the standard deviation was 0.05, and the mean was 0.25 and 0.75 to ensure the separation of the clusters.

Table 1 Generated datasets with number of clusters, total number of instances and the size in MB

Dataset	Clusters	Instances	Size (MB)
5–1k	5	5000	1.00
5–2k	5	10,000	2.00
5–5k	5	25,000	5.00
5–10k	5	50,000	10.00
5–100k	5	500,000	100.00
5–500k	5	2,500,000	501.00
5–1M	5	5,000,000	1003.52
7–1k	7	7000	1.40
7–2k	7	14,000	2.80
7–5k	7	35,000	7.00
7–10k	7	70,000	14.00
7–100k	7	700,000	140.00
7–500k	7	3,500,000	703.00
7–1M	7	7,000,000	1402.88
9–1k	9	9000	1.81
9–2k	9	18,000	3.62
9–5k	9	45,000	9.03
9–10k	9	90,000	18.00
9–100k	9	900,000	181.00
9–500k	9	4,500,000	903.00
9–1M	9	9,000,000	1802.24
11–1k	11	11,000	2.21
11–2k	11	22,000	4.41
11–5k	11	55,000	11.05
11–10k	11	110,000	22.10
11–100k	11	1,100,000	221.00
11–500k	11	5,500,000	1095.68
11–1M	11	11,000,000	2211.84

Table 2 Distance of traditional CVIs to the optimal solution by dataset

	Silhouette	Dunn	David–Bouldin	Calinski–Harabasz	Δ	W	β
5–1k	**0**	**0**	**0**	**0**	**0**	**0**	**0**
5–2k	**0**	**0**	**0**	**0**	**0**	**0**	**0**
5–5k	**0**	**0**	**0**	**0**	**0**	**0**	**0**
5–10k	**0**	**0**	**0**	**0**	**0**	**0**	**0**
7–1k	1	**0**	1	1	1	1	1
7–2k	**0**	**0**	**0**	**0**	**0**	**0**	**0**
7–5k	1	1	1	1	2	2	2
7–10k	2	**0**	2	2	3	2	3
9–1k	2	1	1	2	2	2	2
9–2k	1	2	1	1	1	2	1
9–5k	**0**	1	**0**	1	**0**	**0**	**0**
9–10k	**0**	2	**0**	2	2	3	3
11–1k	1	**0**	2	2	**0**	3	3
11–2k	**0**	**0**	**0**	**0**	**0**	**0**	**0**
11–5k	**0**	2	**0**	**0**	4	**0**	**0**
11–10k	2	2	2	2	2	2	3
Total	10	11	10	14	17	16	17

The hits results are highlighted in bold

4.2 Results

This section is divided in two subsections. Section 4.3 contains the results for the traditional CVIs and Sect. 4.4 shows the results for BD-CVIs. Each subsection includes the effectiveness results and the computational cost. Section 4.3 includes a statistical analysis to compare the effectiveness among CVIs. Section 4.4 provides a statistical analysis to compare the execution time among BD-CVIs.

After executing the CVIs, the effectiveness and execution time of each index are measured. The effectiveness of the indices is calculated by the absolute value of the difference to the optimal solution. Thus, an index with 0 value is considered that it correctly predicts the optimal number of clusters, whereas an index with 2 means that predicts two number above or below the optimal solution. The goodness of fit of the indices is given by the sum of the absolute values of the differences between the real value and the estimated. Therefore, the lower the value, the better the index. The statistical analysis was carried out using the open-source platform StatService [28].

4.3 Results of traditional CVIs

4.3.1 Effectiveness

Table 2 shows the distance to the optimal number of clusters given by the CVIs on each dataset. The datasets were chosen until execution time was under 86,400 s (1 day). The correctly predicted clusters are highlighted in bold, and the last row of the table is the total of distances of each CVI. The lower the value is, the better the CVI is.

Silhouette, Dunn and David–Bouldin obtained the best results since they had the lowest total of distances. The worst results (highest distances) were obtained by Maximum Cluster Diameter (Δ), Average Within-Cluster Distance (W) and Average Between-Cluster Distance (β).

Figure 3a, b shows graphically the number of cluster by Silhouette, Dunn, Davies–Bouldin and Calinski–Harabasz. Δ, W and β were not included in these figures because their results were not so positive. The optimal results are represented by big red dots, so each CVI whose point is on it means that correctly predicted the optimal number of clusters. Datasets with 5 and 7 clusters are included in Fig. 3a and datasets with 9 and 11 clusters are in Fig. 3b.

CVIs obtained very good results in Fig. 3a. Almost all the represented CVIs correctly predicted the optimal solution. Dunn correctly predicted all the datasets except 7–5k. However, Dunn was the only one CVI that set the optimal number of clusters in two datasets in Fig. 3b. 11–10k, 9–1k and 9–2k number of clusters were not estimated by any of the CVI in this figure.

The results of the CVIs do not directly depend on the number of instances of the dataset. The results may be influenced by the number of clusters of the dataset. The lesser number of clusters have the dataset, the greater is the ratio of correct predictions. The optimal number of clusters for datasets with 5 clusters were correctly predicted by all the indices. The optimal number of clusters for datasets with 7 clusters was the most difficult to predict.

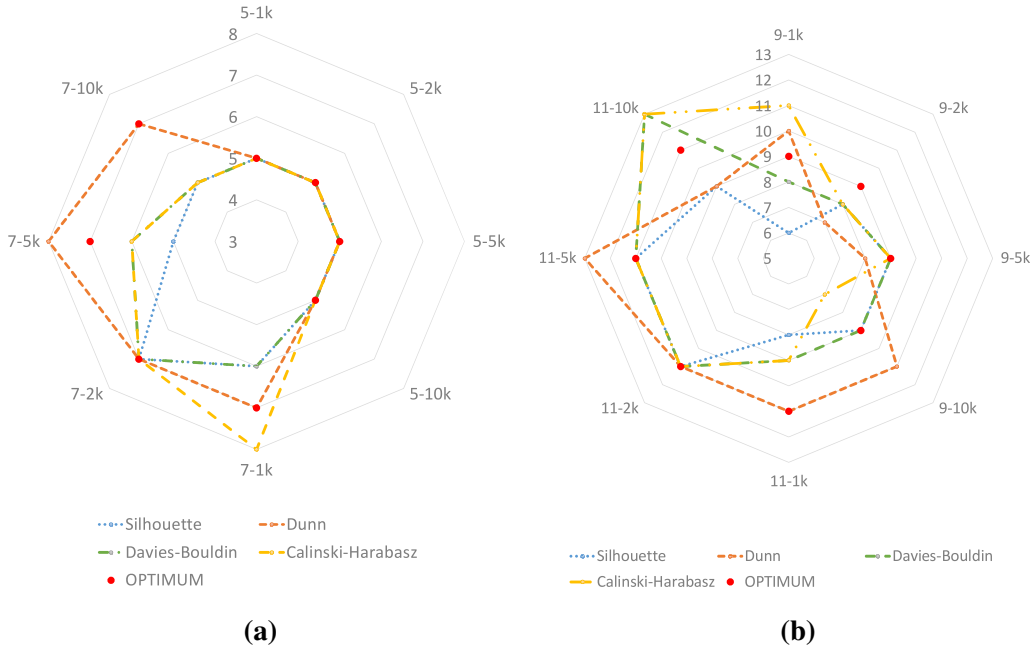


Fig. 3 Results of traditional CVI Silhouette, Dunn, Davies–Bouldin, Calinski–Harabasz for different datasets. Optimal solution by a red dot. **a** Results for datasets with 5 and 7 clusters. **b** Results for datasets with 9 and 11 clusters (color figure online)

Table 3 Average of elapsed time of CVIs in seconds

	Silhouette	Dunn	David–Bouldin	Calinski–Harabasz	Δ	W	β
5–1*k*	9.34	9.33	0.01	2.00	1.99	1.99	7.31
5–2*k*	40.54	40.55	0.01	8.48	8.55	8.59	31.99
5–5*k*	224.46	224.49	0.02	47.20	47.43	47.65	176.91
5–10*k*	1586.37	1584.48	0.48	346.66	340.48	341.01	1238.27
7–1*k*	18.51	18.51	0.01	3.87	3.92	3.92	14.50
7–2*k*	73.48	73.62	0.01	15.73	15.41	15.44	58.14
7–5*k*	906.27	905.93	0.03	450.55	431.62	453.36	368.85
7–10*k*	34,771.06	34,540.32	0.43	91,003.92	86,179.63	99,349.86	3707.81
9–1*k*	20.96	21.12	0.01	2.45	2.35	2.51	18.24
9–2*k*	281.41	278.27	0.02	137.06	136.44	138.55	137.44
9–5*k*	4823.15	4686.26	0.08	14,269.38	10,143.56	9776.66	1330.02
9–10*k*	–	–	0.49	–	–	–	16,552.37
11–1*k*	35.35	34.26	0.01	4.09	3.79	3.55	29.34
11–2*k*	497.75	495.99	0.02	246.05	248.54	246.80	246.21
11–5*k*	8945.03	9178.18	0.14	23,155.30	20,852.84	21,976.94	2653.13
11–10*k*	–	–	1.13	–	–	–	36,766.93

4.3.2 Execution time

Table 3 shows the execution time in seconds of each traditional CVI. Figure 4 is added for better understanding of Table 3. Time is generally increased with the number of instances of the dataset. The lowest execution time was obtained by Davies–Bouldin with very high difference respect to the other CVIs. Davies–Bouldin lasts 1.13 s for the largest dataset (11–10k), while there were some indices whose execution time was higher than 86,400 s (1 day). Such cases are marked as "–". There is a significant increase in the runtime of traditional CVIs for datasets with more than

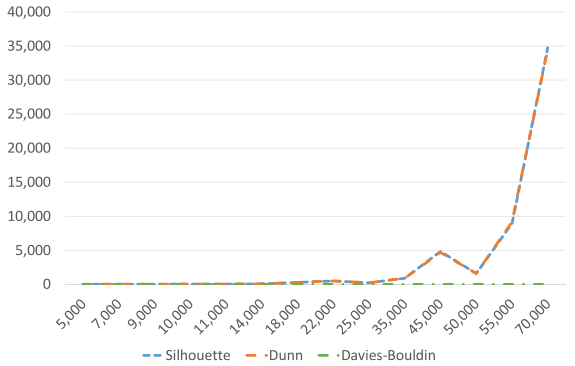


Fig. 4 Representation of traditional CVIs time by the number of instances by dataset

Table 4 Sorted ranking of traditional CVIs for Friedman test

CVI	Ranking
David–Bouldin	3.406
Silhouette	3.531
Dunn	3.656
Calinski–Harabasz	3.969
Δ	4.250
W	4.500
β	4.688

50,000 instances. It can be noted that runtime is four times higher in those datasets even though the growth in the number of instances is less than 50%.

4.3.3 Statistical analysis

After the results generation of the traditional CVIs, a statistical analysis was applied to check if significant differences exist among the effectiveness of the multiple CVIs. The non-parametric Friedman test is shown in Table 4. The highest result for a ranking would be 1, and the worst would be 7. As the ranking shows, Davies–Bouldin was in the first position with 3.406, followed by Silhouette and Dunn with 3.531 and 3.656, respectively.

The statistic for Friedman was 5.0625, distributed according to a Chi-square distribution with 6° of freedom. The p value for Friedman was 0.5358 and higher than 0.05. Therefore, the null hypothesis was accepted that they all behaved in a similar way with a level of significance of $\alpha = 0.05$.

Given the results of Table 3, it makes no sense to perform a statistical test to show that Davies–Bouldin was the fastest CVI.

4.4 Results of BD-CVIs

4.4.1 Effectiveness

BD-CVIs were applied to all datasets from Table 1, including those datasets used in Sect. 4.3. Davies–Bouldin was also included in these experiments for its great results in terms of efficiency in the previous experiments.

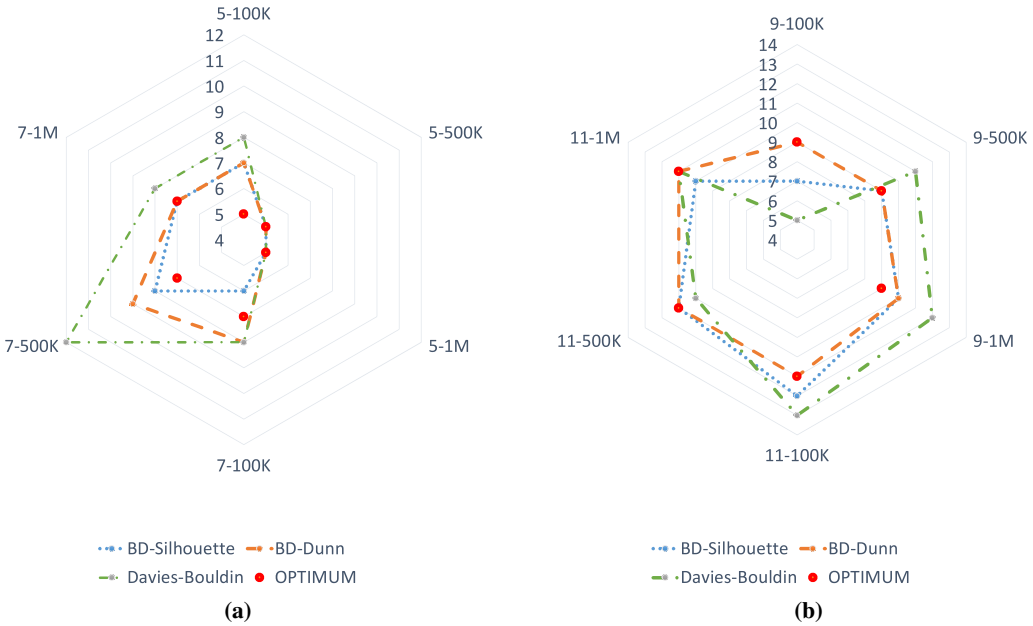


Fig. 5 Results of BD-Silhouette, BD-Dunn and Davies–Bouldin for different datasets. Optimal solution by a red dot. **a** Results for datasets with 5 and 7 clusters. **b** Results for datasets with 9 and 11 clusters (color figure online)

Table 5 Distance of BD-CVIs to the optimal solution by dataset

	BD-Silhouette	BD-Dunn	Davies–Bouldin
5–1*k*	**0**	**0**	**0**
5–2*k*	**0**	**0**	**0**
5–5*k*	**0**	**0**	**0**
5–10*k*	**0**	**0**	**0**
5–100*k*	2	2	3
5–500*k*	**0**	**0**	**0**
5–1*M*	**0**	**0**	**0**
7–1*k*	**0**	**0**	**0**
7–2*k*	**0**	**0**	**0**
7–5*k*	**0**	**0**	**0**
7–10*k*	**0**	**0**	**0**
7–100*k*	1	1	1
7–500*k*	1	2	2
7–1*M*	**0**	**0**	1
9–1*k*	**0**	**0**	**0**
9–2*k*	**0**	**0**	**0**
9–5*k*	**0**	**0**	**0**
9–10*k*	1	3	1
9–100*k*	2	**0**	4
9–500*k*	**0**	**0**	2
9–1*M*	1	1	3
11–1*k*	**0**	**0**	1
11–2*k*	**0**	**0**	**0**
11–5*k*	**0**	**0**	**0**
11–10*k*	1	1	2
11–100*k*	1	**0**	2
11–500*k*	**0**	**0**	1
11–1*M*	1	**0**	**0**
Total	11	10	23

The best results are highlighted in bold

Figure 5a, b shows graphically the results of each BD-CVI by dataset. Red dots highlight the optimal results of each dataset. There were some datasets whose optimal solution was not given by any BD-CVI. However, BD-Silhouette and BD-Dunn correctly predicted most of the datasets; meanwhile, Davies–Bouldin was too far to the optimal like in datasets 7–500*k* or 9–1*M*.

Table 5 shows the distances to the optimal solution of each BD-CVI by dataset. This table shows that the optimal number for datasets with 5 clusters was correctly set by the three indices. Davies–Bouldin did not guess any dataset that BD-Silhouette or BD-Dunn could not. In fact, if BD-Silhouette and BD-Dunn set correctly the optimal number, BD-Davies–Bouldin did it too. There were two cases where BD-Silhouette was the only one BD-CVI that sets the optimal number of clusters correctly.

Table 6 Average of elapsed time of BD-CVIs in seconds

	BD-Silhouette	BD-Dunn	Davies–Bouldin
5–1*k*	0.10	0.06	0.64
5–2*k*	0.11	0.05	0.52
5–5*k*	0.09	0.05	0.63
5–10*k*	0.13	0.10	0.80
5–100*k*	0.22	0.29	0.75
5–500*k*	0.48	0.68	1.26
5–1*M*	0.95	1.67	2.66
7–1*k*	0.11	0.05	0.64
7–2*k*	0.09	0.06	0.61
7–5*k*	0.11	0.08	0.70
7–10*k*	0.12	0.10	0.74
7–100*k*	0.25	0.33	0.66
7–500*k*	0.61	0.94	1.61
7–1*M*	7.15	4.99	6.56
9–1*k*	0.10	0.06	0.80
9–2*k*	0.11	0.06	0.77
9–5*k*	0.10	0.08	0.71
9–10*k*	0.10	0.08	0.83
9–100*k*	0.66	0.85	1.97
9–500*k*	1.72	3.22	3.97
9–1*M*	16.83	9.98	11.42
11–1*k*	0.09	0.06	0.80
11–2*k*	0.12	0.09	0.90
11–5*k*	0.12	0.10	0.98
11–10*k*	0.15	0.20	0.99
11–100*k*	0.28	0.41	1.10
11–500*k*	1.47	2.79	3.69
11–1*M*	25.06	15.23	17.61

4.4.2 Execution time

Table 6 illustrates the total time in seconds after applying BD-CVIs on each dataset. Figure 6 was added to better understand the behavior of Table 6. In datasets where traditional CVIs took more than a day, BD-CVIs took less than 25 s. It is noteworthy that BD-CVIs perform similarly to traditional CVIs. In fact, there is a change in trend of datasets with more than 6 millions instances, as happened in traditional CVIs when the number of instances was 50,000. In the case of BD-CVIs, the runtime had a 400% increase when the number of instances was higher than 6 millions even though the number of instances only has an increment of 100%.

4.4.3 Statistical analysis

Two statistical tests were applied to check the significance in the differences of BD-CVI results, in terms of effectiveness and execution time.

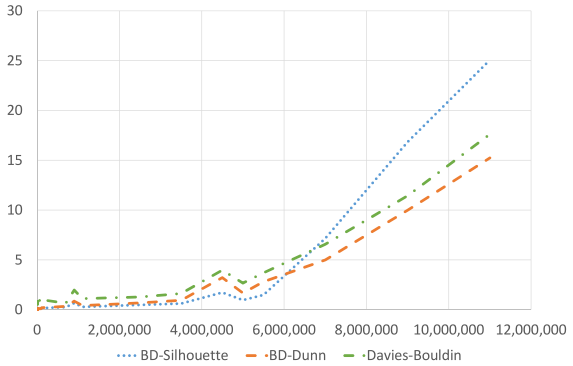


Fig. 6 Representation of BD-CVIs time by the number of instances by dataset

Table 7 Sorted ranking of effectiveness BD-CVI for Aligned Friedman test

BD-CVI	Ranking
BD-Silhouette	35.17
BD-Dunn	36.51
Davies–Bouldin	55.80

Table 8 Post hoc analysis using Holm's procedure and BD-Silhouette as the control algorithm

BD-CVI	p	z	Holm
Davies–Bouldin	0.002	3.164	0.025
BD-Dunn	0.837	0.205	0.050

The effectiveness statistical analysis using Aligned Friedman test is shown in Table 7. Aligned Friedman was used because the test is applied to a dataset with less than 5 features. As the ranking shows, BD-Silhouette was in the first position with 35.17, followed by BD-Dunn with 36.51 and Davies–Bouldin with 55.80 in the last position.

The statistic for Aligned Friedman was 20.45 according to a Chi-square distribution with 2° of freedom. The p value for Aligned Friedman was 0.0 and lower than 0.05. Therefore, the null hypothesis was rejected that they all behaved in a similar way with a level of significance of $\alpha = 0.05$.

Post hoc testing was applied because the null hypothesis that was rejected. Table 8 shows the p values, z value and Holm's α, using BD-Silhouette as the control CVI since it obtained the best ranking. Holm's procedure rejects those hypotheses that have a p value ≤ 0.05.

In execution time statistical analysis, Aligned Friedman test is shown in Table 9. As the ranking shows, BD-Dunn was in the first position with 30.12, followed by BD-Silhouette with 33.44 and Davies–Bouldin with 63.92 in the last position.

The statistic for Friedman was 20.214, distributed according to a Chi-square distribution with 2° of freedom. The p value for Friedman was 0.0 and lower than 0.05. Therefore, the null hypothesis was rejected (that they all behaved in a similar way) with a level of significance of $\alpha = 0.05$.

Table 9 Sorted ranking of execution time BD-CVI for Aligned Friedman test

BD-CVI	Ranking
BD-Dunn	30.12
BD-Silhouette	33.44
Davies–Bouldin	63.92

Table 10 Post hoc analysis using Holm's procedure and BD-Dunn as the control algorithm

BD-CVI	p	z	Holm
Davies–Bouldin	0.000	5.1852	0.025
BD-Silhouette	0.6104	0.5095	0.050

Table 10 shows the p values, z value and Holm's α, using Dunn as the control CVI since it obtained the best ranking. Holm's procedure rejects those hypotheses that have a p value ≤ 0.0083.

4.5 Discussion

Experimental results show that BD-CVIs may be used to provide the optimal number of clusters of large datasets. In this paper, BD-Silhouette and BD-Dunn have achieved better results in lower time than the rest of the indices.

Results show that finding the optimal number of clusters is not a trivial task. There were some datasets that were not correctly solved. The results in this study indicate that Silhouette, Dunn and Davies–Bouldin were the CVIs with the highest success rate. This fact is particularly significant because it helps to construct new CVIs that are suitable to work with Big Data.

This study also found that BD-CVIs had even more difficult to provide the optimal number of cluster of a dataset. The results of this study indicate that there were complex datasets where no BD-CVI correctly predicted the optimal number of clusters. All these support the notion that getting the optimal number of clusters is not a minor task. However, the results of this study show that BD-Silhouette and BD-Dunn are good choices to predict the optimal number of clusters as the results were promising.

In terms of time, traditional indices last so much time compared with BD-CVI. The results of this study indicate that the biggest dataset used with traditional indices last more than a day; however, BD-CVIs in the same dataset lasted less than 1 minute. These observations provide evidence that suggests that the use of traditional indices is very limited due to the size of the datasets.

5 Conclusions

In this paper, two novel CVIs implemented in Spark have been proposed to be applied in datasets considered as Big Data. The proposed indices are based on Silhouette and Dunn indices, but modified and optimized to deal with Big Data.

The experimental study indicates that our Big Data indices are very competitive. We have tested its effectiveness and time execution with datasets of different sizes (different number of clusters and different number of instances). The main achievements obtained are the following:

- Two clustering indices based on traditional Silhouette and Dunn indices.
- BD-Silhouette and BD-Dunn has allowed us to estimate the optimal number of clusters of datasets that may be considered Big Data.
- Computational time of these indices is drastically reduced compared with traditional indices.
- The size of the dataset does not directly influence to the effectiveness of the BD-CVIs.
- The software of this contribution can be found as a spark-package at http://spark-packages.org/package/josemarialuna/clusterIndices.
- The source code of these indices can be found at https://github.com/josemarialuna/ClusterIndices.

As a future work, we intend to include some approaches to other CVIs that also obtained suitable results in their traditional version. Further research is needed to study the outcomes because some of the BD-CVI results were not enough clear. It would also be useful to explore the results of our BD-CVIs with no round-shape clusters datasets. Additionally, it would be also interesting to research the results of BD-CVIs taking into consideration using the inter-cluster distances between the centroids instead of using the global centroid.

Acknowledgements This work has been supported by the Spanish National Research Project TIN2014-55894-C2-1-R. J.M. Luna-Romera holds an FPI scholarship from the Spanish Ministry of Education.

Appendix A: Datasets generation

In this work, we needed suitable datasets to cluster the data and to know in advance how many clusters have them. We have developed an application that generates especially designed datasets with predefined number of clusters. To make sure that the clusters are well formed and separated, the points of the datasets follow a normal distribution with different mean values and a low standard deviation. Datasets are generated introducing as input parameters the following items: the total number of clusters of the dataset, the number of instances per cluster, the number of features of the instances and the standard deviation. As we mention before, feature values of data in clusters follow a normal distribution, and for this purpose, data is randomly generated with the given standard deviation and the mean, that by default is 0.25 or 0.75. With this random generation of points, we ensure that clusters are well separated and it will make easier to be identified by the CVIs.

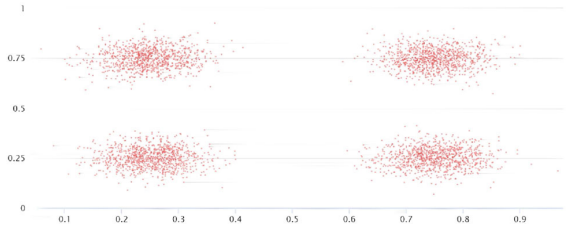


Fig. 7 Generated dataset with 4 clusters and 2 features generated with a mean of 0.25 and 0.75 and a standard deviation of 0.05

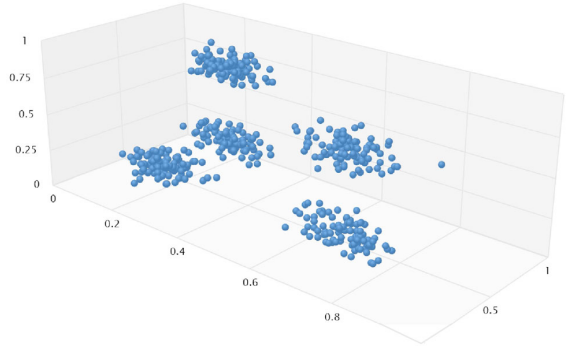


Fig. 8 Generated dataset with 5 clusters and 3 features with a mean of 0.25 and 0.75 and a standard deviation of 0.05

Figures 7 and 8 illustrate two basics example datasets that were generated by the application. Those figures show the distribution of the points in 2D and 3D. These datasets, and those generated for the experiments, were created with an average of 0.25 and 0.75, and a standard deviation of 0.05. Figure 7 corresponds to a dataset with 2 features and 4 clusters using 1000 instances per cluster. As it can be seen, in this figure there are 4 clear groups of points that correspond with the clusters. There are also some points that are not close to a big cloud of points and this is due to data points are randomly generated following a normal distribution a its standard deviation is 0.05.

A similar situation is found in Fig. 8. This figure is a 3D representation of a dataset with 3 features where 5 clusters can be easily identified. Each cluster counts with 100 instances and, as it happened in Fig. 7, there are also some points that are separated from the central cloud of points.

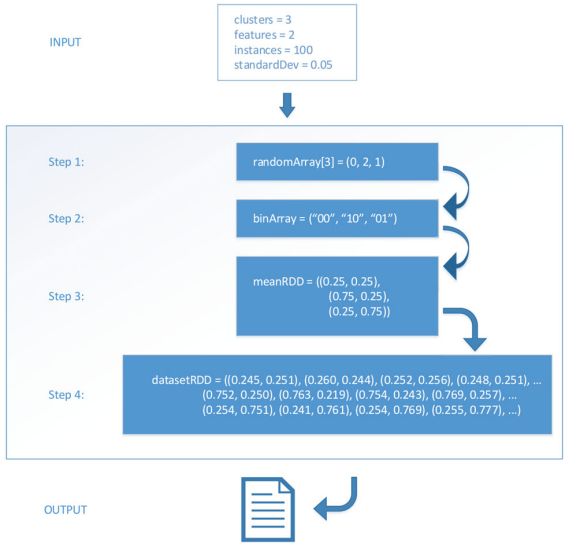


Fig. 9 Flowchart of dataset generator algorithm

Figure 9 shows graphically step by step how the application generates a dataset. The figure shows the input of the algorithm represented by a white box at the top, the application with the steps represented by a darker box, and the output file at the bottom of the figure. In our example the application receives the next input parameters: 3 clusters, 2 features, 100 instances per cluster, and a standard deviation of 0.05.

1. *Step 1* the application receives the input parameters and an array named $randomArray$, whose size is the number of clusters. $randomArray$ is randomly generated with the numbers in the interval $[0, numberofclusters)$ without repetition. In the figure is shown that $randomArray$ has size 3, and the random included numbers are (0, 2, 1).
2. *Step 2* the values of $randomArray$ are parsed to binary, and they are saved into an array named $binArray$. In our example, $randomArray$ was (0, 2, 1), so $binArray$ becomes (00, 10, 01).
3. *Step 3* $meanRDD$ is an RDD object that takes its values from $binArray$. The values of the array are individually taken, and if it is 0, it sets 0.25; or if it is 1, it sets 0.75. In our example, "00" becomes (0.25, 0.25), "10" becomes (0.75, 0.25), and "01" becomes (0.25, 0.75).
4. *Step 4* on this step, the values of each data object in the dataset are generated and saved into an RDD object. It generates $instances$ value random numbers following a normal distribution with the standard deviation given as input parameter ($standardDev$), and the mean is set by the value of $meanRDD$. Each value of $meanRDD$ will be the data objects of each cluster. In our example, the application will generate 100 data objects with and a standard deviation of 0.05, and a mean of 0.25 for the first feature, and a 0.25 for the second feature ((0.245, 0.251), (0.260, 0.244), (0.252, 0.256)…). The data objects of the second clusters take (0.75, 0.25) as the values for the mean and generate the following data objects: ((0.752, 0.250), (0.763, 0.219), (0.754, 0.243)…). And the third cluster has the following data objects: ((0.254, 0.751), (0.241, 0.761), (0.254, 0.769)…).
5. Once $datasetRDD$ is built, the application saves the data into an output file.

The source code of this application can be found at [24].

References

1. Abdi, A., Hassanzadeh, Y., Ouarda, T.: Regional frequency analysis using Growing Neural Gas network. J. Hydrol. **550**, 92–102 (2017)
2. Alok, A., Saha, S., Ekbal, A.: Semi-supervised clustering for gene-expression data in multiobjective optimization framework. Int. J. Mach. Learn. Cybern. **8**(2), 421–439 (2017)
3. Berikov, V., Pestunov, I.: Ensemble clustering based on weighted co-association matrices: error bound and convergence properties. Pattern Recognit. **63**, 427–436 (2017)
4. Boone, C., Skipper, J., Hazen, B.: A framework for investigating the role of big data in service parts management. J. Clean. Prod. **153**, 687–691 (2017)
5. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. Theory Methods **3**(1), 1–27 (1974)
6. Chen, W.-Y., Song, Y., Bai, H., Lin, C.-J., Chang, E.Y.: Parallel Spectral Clustering in Distributed Systems. IEEE Trans. Pattern Anal. Mach. Intell. **33**(3), 568–586 (2011)
7. Daki, H., El Hannani, A., Aqqal, A., Haidine, A., Dahbi, A.: Big Data management in smart grid: concepts, requirements and implementation. J. Big Data **4**(1), 13 (2017)
8. Davies, D., Bouldin, D.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI–1**(2), 224–227 (1979)
9. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
10. Dubes, R., Jain, A.K.: Clustering techniques: the user's dilemma. Pattern Recognit. **8**(4), 247–260 (1976)
11. Dunn, J.: Well-separated clusters and optimal fuzzy partitions. J. Cybern. **4**(1), 95–104 (1974)
12. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A .Y., Foufou, S., Bouras, A.: A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Trans. Emerg. Top. Comput. **2**(3), 267–279 (2014)
13. Gallos, L., Korczyński, M., Fefferman, N.: Anomaly detection through information sharing under different topologies. Eurasip J. Inf. Secur. **1**, 2017 (2017)
14. Ghemawat, S., Gobioff, H., Leung, S.-T.: The Google File System, vol. 37, pp. 29–43. ACM Press, New York, USA (2003) **(cited By 2613)**
15. Han, J., Kamber, M., Pei, J.: Cluster analysis: basic concepts and methods. In: Data Mining: Concepts and Techniques, pp. 443–495. Elsevier, USA (2012)
16. Hennig, C., Liao, T.: How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. J. R. Stat. Soc. Ser. C Appl. Stat. **62**(3), 309–369 (2013)
17. Holmes, G., Donkin, A., Witten, I.: WEKA: a machine learning workbench. In: Proceedings of ANZIIS '94—Australian New

Zealnd Intelligent Information Systems Conference, Number JANUARY 1994, pp. 357–361. (1994)
18. Jacques, J., Preda, C.: Functional data clustering: a survey. Adv. Data Anal. Classif. **8**(3), 231–255 (2014)
19. Jain, A .K.: Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. **31**(8), 651–666 (2010)
20. Jerome, R .B., ätönen, K .H.: Anomaly detection and classification using a metric for determining the significance of failures. Neural Comput. Appl. **28**(6), 1265–1275 (2017)
21. Jinyin, C., Xiang, L., Haibing, Z., Xintong, B.: A novel cluster center fast determination clustering algorithm. Appl. Soft Comput. **57**, 539–555 (2017)
22. Kim, J., Lee, W., Song, J .J., Lee, S.-B.: Optimized combinatorial clustering for stochastic processes. Clust. Comput. **20**(2), 1135–1148 (2017)
23. Lord, E., Willems, M., Lapointe, F.-J., Makarenkov, V.: Using the stability of objects to determine the number of clusters in datasets. Inf. Sci. **393**, 29–46 (2017)
24. Luna-Romera, J.M.: Clustering Synthetic Big Datasets Generator. https://github.com/josemarialuna/CreateRandomDataset (2017). Accessed 20 July 2017
25. Mazinan, A.: On cluster validity indices with its application to interleaved radar pulse separation through fuzzy-based representation. Evol. Syst. **7**(4), 243–254 (2016)
26. Miller, Z., Dickinson, B., Deitrick, W., Hu, W., Wang, A.H.: Twitter spammer detection using data stream clustering. Inf. Sci. **260**, 64–73 (2014)
27. Mohammed, A.J., Yusof, Y., Husni, H.: Fireflyclust: an automated hierarchical text clustering approach. Jurnal Teknologi, **79**(5), 11–22 (2017)
28. Parejo, J.A., Garcia, J., Ruiz-Cortes, A., Riquelme, J.C.: Statservice: Herramienta de análisis estadistico como soportepara la investigacion con metaheuristicas. In: Actas del VIII Congreso Expañol sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados. Albacete, España (2012)
29. Perez-Chacon, R., Talavera-Llames, R., Martinez-Alvarez, F., Troncoso A.: Finding Electric Energy Consumption Patterns in Big Time Series Data. In: Omatu, S., et al. (eds.) Distributed Computing and Artificial Intelligence, 13th International Conference. Advances in Intelligent Systems and Computing, vol. 474, pp. 231–238. Springer, Cham (2016)
30. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**(C), 53–65 (1987)
31. Rumson, A .G., Hallett, S .H., Brewer, T .R.: Coastal risk adaptation: the potential role of accessible geospatial Big Data. Mar. Policy **83**, 100–110, (2017)
32. Sagi, T., Gal, A., Barkol, O., Bergman, R., Avram, A.: Multi-source uncertain entity resolution: transforming holocaust victim reports into people. Inf. Syst. **65**, 124–136 (2017)
33. Sevilla-Villanueva, B., Gibert, K., ànchez-Marrè, M.S.: Using CVI for Understanding Class Topology in Unsupervised Scenarios, pp. 135–149. Springer, Cham (2016)
34. Spark, A.: Apache Spark, Lightning-Fast Cluster Computing. https://spark.apache.org/ (2017). Accessed 20 June 2017
35. Spark, A.: MLlib is Apache Spark's Scalable Machine Learning Library. https://spark.apache.org/mllib/ (2017). Accessed 20 June 2017
36. Tong, Q., Li, X., Yuan, B.: A highly scalable clustering scheme using boundary information. Pattern Recognit. Lett. **89**, 1–7 (2017)
37. Yang, M., Mei, H., Huang, D.: An effective detection of satellite images via k-means clustering on hadoop system. Int. J. Innov. Comput. Inf. Control **13**(3), 1037–1046 (2017)
38. Zaharia, M.. Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., Franklin, M.J., Shenker, S., Stoica, I.: Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Presented as Part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12), pp. 15–28, San Jose, CA, USENIX (2012)
39. Zhang, Q., Yang, L.T., Chen, Z., Li, P.: High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT. Inf. Fusion **39**, 72–80 (2018)
40. Zhang, R., Xu, C., Duan, Z.: Novel antigenic shift in HA sequences of H1N1 viruses detected by big data analysis. Infect. Genet. Evol. **51**, 138–142 (2017)

Capítulo 5

External clustering validity index based on chi-squared statistical test

Resumen

Este artículo de investigación presenta un nuevo índice de validación externo de clustering basado en el test estadístico de independencia de variables cualitativas chi cuadrado que hemos denominado Chi Index. Los índices de validación externos son aquellos que miden la bondad de un clustering basándose en algún atributo externo a los incluidos a la hora de hacer el clustering, de manera que miden la calidad del clustering en función de una etiqueta o clase. Normalmente los índices externos de la literatura muestran su resultado basándose en representaciones gráficas cuyas interpretaciones pueden llevar a error. Chi Index presenta el resultado de la solución óptima de clustering sin necesidad de ser interpretado. Además, al estar basado en chi cuadrado mide la relación existente entre la distribución de los puntos por los *clusters* y por las clases, de manera que un buen resultado de clustering será aquel cuyas instancias estén separadas por clases y *clusters* al mismo tiempo. La experimentación se ha llevado a cabo en dos partes: en la primera se prueba que el índice cumple una serie de propiedades únicas en los índices externos en diferentes datasets sintéticos; en la segunda se prueba la eficiencia del índice comparando sus resultados en 47 datasets públicos frente a otros 15 CVIs de la literatura y tomando como resultado los clustering de 3 métodos diferentes. Los resultados de este artículo muestran que chi index es significativamente mejor al resto de índices de la literatura, y además ofrece el resultado de manera exacta sin necesidad de que sea interpretado.

- Estado: Publicado en Information Sciences (Elsevier) (2019), Volumen: 487, 1-17

- Índice de Impacto (JCR 2018): 5.524
- Área de Conocimiento:
  - Computer Science, Information Systems. Ranking 12/105 - Q1

Information Sciences 487 (2019) 1–17

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

External clustering validity index based on chi-squared statistical test

José María Luna-Romera*, María Martínez-Ballesteros, Jorge García-Gutiérrez, José C. Riquelme

Department of Computer Languages and Systems, ETSII, University of Seville, Spain

ARTICLE INFO

Article history:
Received 2 July 2018
Revised 15 February 2019
Accepted 17 February 2019
Available online 18 February 2019

Keywords:
Clustering analysis
External validity indices
Comparing clusters
Big data

ABSTRACT

Clustering is one of the most commonly used techniques in data mining. Its main goal is to group objects into clusters so that each group contains objects that are more similar to each other than to objects in other clusters. The evaluation of a clustering solution is a task carried out through the application of validity indices. These indices measure the quality of the solution and can be classified as either internal that calculate the quality of the solution through the data of the clusters, or as external indices that measure the quality by means of external information such as the class. Generally, indices from the literature determine their optimal result through graphical representation, whose results could be imprecisely interpreted. The aim of this paper is to present a new external validity index based on the chi-squared statistical test named Chi Index, which presents accurate results that require no further interpretation. Chi Index was analyzed using the clustering results of 3 clustering methods in 47 public datasets. Results indicate a better hit rate and a lower percentage of error against 15 external validity indices from the literature.

1. Introduction

Clustering is one of the many techniques in data mining. Its goal is to partition unlabelled data into clusters where instances within the same cluster are similar and instances grouped in other clusters are dissimilar to said clusters [1]. This technique has been applied in many fields, such as biological sciences [2], medicine [3], energy [4], chemical [5].

There are numerous clustering methods, and in general, each method produces a different clustering solution. In certain cases, the same method with different parameters could result in different solutions. The evaluation of the results is one of the most important issues in cluster analysis. Measuring the quality of a clustering solution is as important as the clustering method itself [6]. There exist clustering validity indices (CVI) that measure the quality of the solution, and these CVIs have commonly been used in the literature [7–13].

These measures could be classified into either internal or external CVIs. Internal CVIs are based on how the instances are distributed across the clusters by using the data by itself. When there is no external information, these kinds of indices present the only option available for the evaluation of the clustering solution because they depend on certain properties of the results, such as the compactness of the clusters or the separation between them. Compactness of clusters could be

* Corresponding author.
E-mail addresses: jmluna@us.es (J.M. Luna-Romera), mariamartinez@us.es (M. Martínez-Ballesteros), jorgarcia@us.es (J. García-Gutiérrez), riquelme@us.es (J.C. Riquelme).

https://doi.org/10.1016/j.ins.2019.02.046
0020-0255/

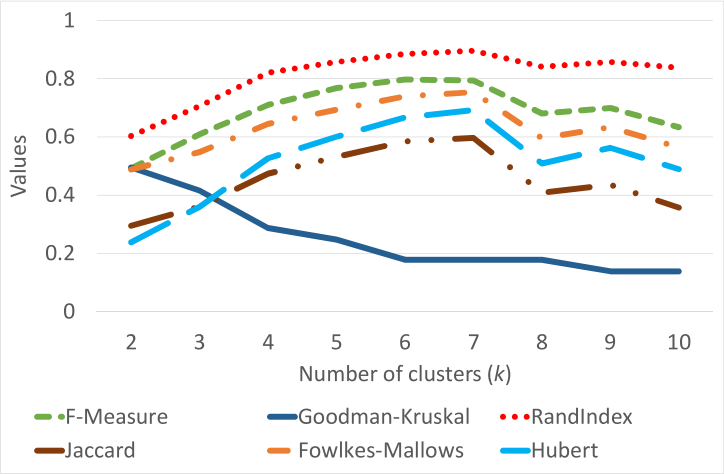


Fig. 1. Results of the CVIs from the literature for $k = 2$ to 10 number of clusters for zoo dataset whose optimal number of clusters is 7.

defined as the mean distance of separation between the instances within a cluster. Separation by itself is defined as the distance between the instances of different clusters. These indices seek a high level of compactness within each cluster and a considerable gap between clusters [14].

On the other hand, external indices use external information, such as class labels, to measure the quality of a clustering solution. These kinds of indices verify the quality of the clustering result by comparing it with the ground truth partition. In this case, the indices know in advance the optimal number of clusters for a dataset since ground truth holds this information [15]. This paper focuses on these external CVIs. Generally, CVIs from the literature determine their optimal result with a local minimum, a local maximum, or by following the elbow method [16–18], and the results could be imprecisely interpreted.

The purpose of this paper is to present an innovative external CVI based on the chi-squared statistical test, henceforth named Chi Index, which presents the results accurately without the need for interpretation. The effectiveness of the new index has been compared with 15 indices from the literature using 47 public datasets and 3 clustering methods from Spark MLlib [19] which made it possible to use this index in big data environments.

The remainder of this paper is organized as follows. Section 2 discusses the literature of external CVIs. In Section 3, the proposed new index is defined. Section 4.3 presents the experimental setup, the methodology followed and the results. The paper ends with the conclusions and suggested future work in Section 5.

2. External indices

An external index evaluates a clustering result C by comparing it against the ground truth partition G. A taxonomy of external indices could be established that depends on the criterion of how the clustering result and the ground truth partition are compared [20]. These indices can be classified into *set matching*, *pair-counting*, and *information theory*.

- *Set matching* is the category which assumes that the instance label of every cluster has corresponding instances in said cluster. Indices from the literature based on *set matching* include those known as *purity* [21], *F-measure* [22], *Criterion H* [23], *CSI* [24], *PSI* [20], and *Goodman–Kruskal* [25].
- The criterion known as *pair-counting* is based on the comparison between the number of instances with the same label and the cluster result. This category includes the *Rand index* [26], the *adjusted Rand index* [27], *Jaccard* [28], *Fowlkes–Mallows* [29], *Hubert Statistic* [30], and *Minkowski score* [31].
- Indices based on *information theory*, such as *entropy* [21], *variation of information* [32], and *mutual information* [33], have also been applied in the literature.

A list of the equations of these indices is given in Table 1. As mentioned above, the results that show these indices need to be interpreted since each index indicates the optimal number following the rules of the local maximum, the local minimum, or the “elbow method”. Figs. 1 and 2 illustrate two examples of the results for the CVIs from the literature for zoo and gesture datasets from the UCI repository whose optimal number of clusters is 7 and 5, respectively. In Fig. 1, it could be said that the CVIs follow a pattern, whereby the majority indicate point out the optimal number of clusters to be 7 with a local maximum, although Goodman–Kruskal indicates the optimal by following the elbow method. This figure shows that most of the CVIs also have a local maximum at 9, and this could be misleading in the cases when the optimal number of clusters remains unknown in advance. Fig. 2 corresponds to a dataset whose optimal number of clusters is 4; however, no index clearly shows the solution. The F-Measure, Jaccard, Fowlkes–Mallows, and Hubert indices, which indicate the optimal number with maximum values, all have a local maximum not only at 5 but also at 8. Furthermore, the

Table 1
Equations of external clustering validity indices from the literature equations.

Preliminaries	
Total elements in the dataset	n
Elements in cluster i in class j	n_{ij}
Total elements in cluster i	$n_{i.}$
Total elements in class j	$n_{.j}$
Rate of the cell ij	$p_{ij} = \frac{n_{ij}}{n}$
Rate of the row i	$p_{i.} = \frac{n_i}{n}$
Rate of the colum j	$p_j = \frac{n_j}{n}$
Set matching	
Purity [42]	$P = \sum_i p_i (max_j \frac{p_{ij}}{p_i})$
F-Measure [20]	$FM = \sum_j p_j max_i (2 \frac{\frac{p_{ij}}{p_i} \frac{p_{ij}}{p_j}}{\frac{p_{ij}}{p_i} + \frac{p_{ij}}{p_j}})$
Goodman-Kruskal [12]	$GK = \sum_i p_i (1 - max_j \frac{p_{ij}}{p_i})$
Criterion H [25]	$CH = 1 - \frac{1}{n} \max \sum_{i=1}^{k} n_{ij}$
CSI [10]	$CSI = \frac{\sum_{i=1}^{k} n_{ij} + \sum_{i=1}^{k'} n_{ji}}{2n}$
PSI [30]	$PSI = \begin{cases} \frac{S - E(S)}{\max(k,k') - E(S)} & S \geq E(S), \max(k,k') > 1 \\ 0 & S < E(S) \\ 1 & K = K' = 1 \end{cases}$
Pair-counting	
Rand index [29]	$RI = [\binom{n}{2} - \sum_i \binom{n_{i.}}{2} - \sum_j \binom{n_{.j}}{2} + 2\sum_{ij} \binom{n_{ij}}{2}] / \binom{n}{2}$
Adjusted rand index [36]	$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}] / \binom{n}{2}}{\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}] / 2 - [\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}] / \binom{n}{2}}$
Jaccard [33]	$J = \frac{\sum_{ij} \binom{n_{ij}}{2}}{\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} - \sum_{ij} \binom{n_{ij}}{2}}$
Fowlkes and Mallows [9]	$FM = \frac{\sum_{ij} \binom{n_{ij}}{2}}{\sqrt{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}}$
Hubert Statistic [17]	$H = \frac{\binom{n}{2} \sum_{ij} \binom{n_{ij}}{2} - \sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\sqrt{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} [\binom{n}{2} - \sum_i \binom{n_{i.}}{2}] [\binom{n}{2} - \sum_j \binom{n_{.j}}{2}]}}$
Minkowski Score [3]	$MS = \frac{\sqrt{\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} - 2\sum_{ij} \binom{n_{ij}}{2}}}{\sqrt{\sum_j \binom{n_{.j}}{2}}}$
Information Theory	
Entropy [42]	$E = -\sum_i p_i (\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$
Variation of Information [24]	$VI = -\sum_i p_i \log p_i - \sum_j p_j \log p_j - 2\sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$
Mutual Information [2]	$MI = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$

Table 2
Three different distribution examples with 3 classes (A, B, C) and 3 clusters (1, 2, 3).

(a) Contingency table where chi-squared is 0.				(b) Contingency table where chi-squared reaches its maximum value.				(c) Contingency table in which the distribution of the instances could be found on a real scenario.			
Cluster	A	B	C	Cluster	A	B	C	Cluster	A	B	C
1	2	2	2	1	6	0	0	1	3	3	0
2	1	1	1	2	0	0	3	2	0	3	0
3	3	3	3	3	0	9	0	3	0	0	9

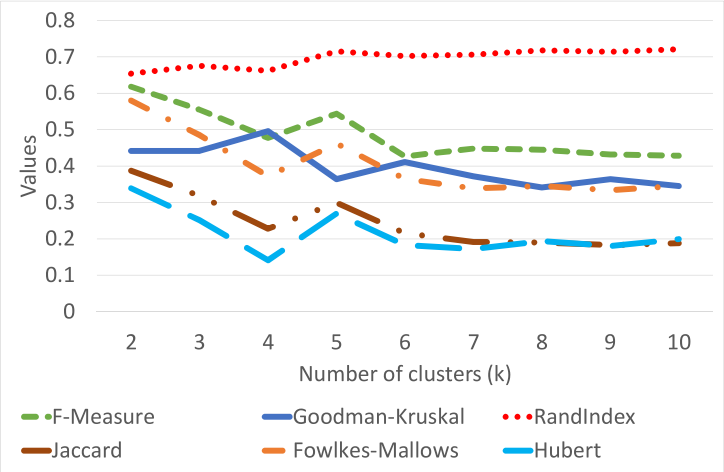


Fig. 2. Results of the CVIs from the literature for $k = 2$ to 10 number of clusters for knowledge dataset whose optimal number of clusters is 4.

Goodman–Kruskal index, which reaches its optimal number of clusters at the minimum value, has a low local minimum at 5 and at 8. Additionally, the Rand Index, following the elbow method, marks the optimal number at 5. In summary, CVIs indices can be misleading due to the interpretation of its results.

In recent years, several studies that propose new external indices for clustering validation have been published in the literature.

A new *pair-counting* index, which is based on an intuitive probabilistic approach, is employed to compare solutions that may have a certain degree of overlap in [34]. This index was tested using four artificial datasets with 6 classes and 4 real datasets from the UCI repository [35].

A new index was also presented in [36], but in this case, it is based on Max-Min distance between data points and prior information. This external index could be classified in the category of *set matching*. The performance of this index was compared with *set matching* and *pair-counting* indices using 6 artificial datasets and two real datasets also from the UCI repository.

The authors of the work presented in [37], proposed a new index based on an ensemble of supervised classifiers. We may classify this index as a *pair-counting* index. Fifty real datasets from the UCI repository were used for the experiments and the results were compared with several internal indices.

A new *pair-counting* index for analytical comparisons was presented in [20]. It applies a correction for chance and normalizes for each cluster separately. The experiments were carried out with artificial datasets with 3 classes and 6000 instances in each dataset. This new index obtained better results than other external CVIs such as purity, adjusted rand index, and mutual information.

In [10], other authors suggested a new *set-matching* index based on the conception of a degree of freedom that measures the decision interval between two classes. This index measures the quality of the clustering by comparing it with internal and *set matching* external indices. Fourteen real datasets were used to test the performance of the index.

Most of these clustering validation techniques are verified by comparing the clustering results with CVIs from the literature and by using synthetic datasets. This work strives to provide a reliable, and accurate CVI based on the chi-squared statistical test as the basis for clustering analysis.

3. Proposed external clustering validity index based on the chi-squared test

3.1. Chi-squared

The Pearson chi-squared statistical test is a method that determines whether there exists a significant difference between the expected values and the observed values in a distribution between two variables [38]. The following equation is applied to verify this correlation:

$$\chi^2 = \sum_{i}^{r} \sum_{j}^{c} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{1}$$

where r is the number of rows, c is the number of columns, n_{ij} is the observed value and E_i is the expected value. E_i is given by

$$E_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} \tag{2}$$

Table 3
Contingency tables of Table 2c expressed in terms of relative frequencies.

(a) By relative frequencies per row.

#	A	B	C	
1	50%	50%	0%	100%
2	0%	100%	0%	100%
3	0%	0%	100%	100%

(b) By relative frequencies per column.

#	A	B	C
1	100%	50%	0%
2	0%	50%	0%
3	0%	0%	100%
	100%	100%	100%

where n is the total number of instances.

The χ^2 value is employed to determine the suitability of the value through the significant interval. In this way, χ^2 approaches to zero when the observed value resembles the expected value. Therefore, if the observed values are similar to the mean, χ^2 indicates that there is no dependence between the two variables that are being analysed.

3.2. Motivation

External validity clustering indices measure the quality of the clustering result by focusing on a ground truth. Our Chi Index may be considered a set-matching measure since it matches the clusters, and measures the similarity between the clustering and the ground truth, which is given by the maximum value that Chi Index could reach. In addition, the Chi Index is normalized in order to be influenced neither by the number of clusters nor the number of classes. The strategy of the Chi Index is, in general terms, to set the instances of the same class in separate clusters in such a way that the instances which belong to the same class are grouped together as much as possible. In addition, the Chi Index aims to define each cluster by a single class as far as possible. Therefore, the Chi Index looks for the clustering solution that, on the one hand, separates the classes into clusters, and, on the other hand, splits the clusters so that each one can be identified by a class.

The chi-squared test measures the difference between the expected frequencies and the observed frequencies in a distribution. The lower the chi-squared value, the more similar the expected values are to the observed values, that is, if the observed values of the distribution are closer to the mean, then the chi-squared value approaches zero.

Table 2 presents 3 contingency matrices for a distribution with 3 classes (A, B, C) and 3 clusters (1, 2, 3). The values in Table 2a are the same for all the clusters within the classes; in this case, the chi-squared value is 0. The Chi Index seeks exactly the opposite scenario, where the clusters are formed by only one class and where each class is only presented in one cluster, as illustrated in Table 2b. Table 2c presents a distribution where cluster 1 is formed of instances of classes A and B, cluster 2 is composed of instances of only class B, and cluster 3 is consisted of instances from class C.

In order to ensure that each class is only presented in one cluster and each cluster has only one class, the values of the contingency matrix have to be expressed in relative terms. To this end, the absolute frequency contingency table has to be transformed into 2 contingency matrices, one for the relative frequencies per row, and the other for the relative frequencies per column. Hence, in the first contingency matrix, the sum of the rows is 100%, and in the second contingency matrix, the sum of the columns is also 100%.

Taking Table 2c as an example, Table 3a and b are built transforming the absolute frequencies into relative frequencies. As mentioned before, the tables are expressed in relative terms to the total of rows and columns.

In this way, Table 3a indicates that cluster 1 is evenly split between classes A and B, cluster 2 is composed of instances from class 2, and cluster 3 has instances only from class 3. Alternatively, Table 3b shows that the instances from class A are only in the cluster 1, the instances from class B are evenly split between clusters 1 and 2, and the instances from class C are only in cluster 3.

In addition, the Chi Index has an accurate result that needs no interpretation. If we analysed the results for the Chi Index iterating over the number of clusters k, we would obtain two curves, one for each contingency matrix. In general, the clusters tend to become more specialized as the number of clusters increases, that is, there is a higher percentage of points of the same class in each cluster which will increase the chi-squared value for the matrix per row. On the other hand, when the records of each class are distributed across a greater number of clusters, then the value of the chi-square per column will tend to decrease. Our goal is to simultaneously maximize both values by encouraging their tendency to diverge. The first value where both series are cut off (or the distance between them is minimized as we cannot be sure wheter they will be crossed) sets the optimal number of clusters in our proposal. Henceforth, the Chi Index identifies the optimal solution as the minimum difference between the chi-squared values of the curves, thereby rendering it unnecessary to interpret the result thanks to its accuracy.

3.3. Chi Index toy example

Fig. 3 illustrates the spatial distribution of the instances of our toy example dataset with 24 instances and 3 classes. Each dot represents an instance and its colour defines the class to which it belongs.

Before applying a clustering method to this dataset, the number of clusters has to be previously determined. Fig. 4 shows the clustering solution from $k = 2$ to 4. It is difficult to determine which clustering solution is the best at a glance.

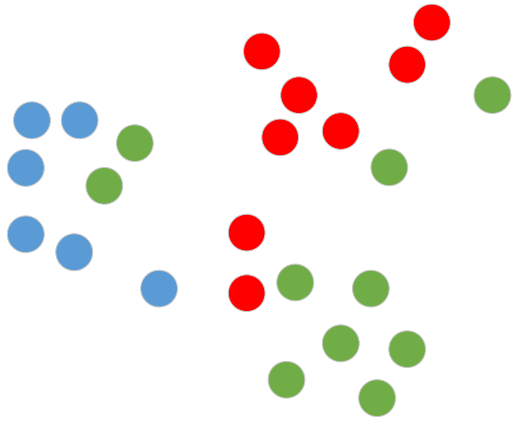
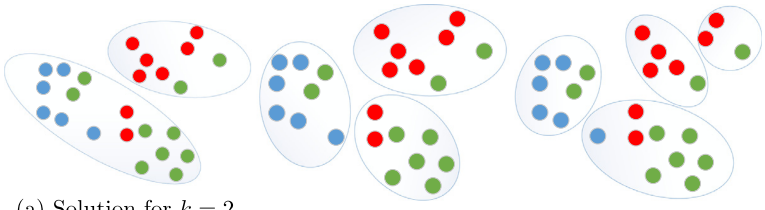


Fig. 3. Representation of the instance distribution of the toy example.



(a) Solution for $k = 2$. (b) Solution for $k = 3$. (c) Solution for $k = 4$.

Fig. 4. Clustering solution representation for $k = 2$ to 4.

To this end, an index that measures the quality of each clustering solution and selects the best one is required. The Chi Index measures the quality of the clustering based on the chi-squared test.

If we focus on the toy example, Fig. 4a represents the clustering solution for $k = 2$. This figure shows that cluster 1 has 2 instances from the red class, 8 green instances, and 6 blue instances, while cluster 2 has 6 red instances, 2 greens instances, and none from blue class. This information is shown in a contingency table in Table 4a, where the clusters are represented by rows, and the classes red, green, and blue are R, G, and B respectively. This table could be analysed in two ways: by rows, where we can conclude that cluster 1 is mainly composed of green instances, but it also has red and blue instances. However, cluster 2 is only composed of red and green instances.; by columns, where blue instances are only in cluster 1, red and green instances are distributed in both clusters.

This analysis is illustrated in Table 4d, where the relative frequency of the instances are expressed in relation to the total of rows (left-side) and columns (right-side).

A complete representation of each clustering solutions from $k = 2$ to 4 is presented in Table 4 with a pair of tables: the contingency table with the absolute frequency, and the contingency tables with relative values by rows and by columns.

Once we have the contingency tables with the relative values, we need to obtain the chi-square value of these tables for each iteration. In our toy example, the Chi Index has been calculated for the clustering solutions with $k = 2$ to 4. The goal is to maximize the values of the Chi Index in both tables and minimize the difference between them. Thus, the Chi Index result will ensure that the observed and expected values differ as much as possible, thereby keeping the solution with the highest percentage of classes in each cluster. Eqs. (3) and (4) detail how the chi square value by row and by column are calculated respectively for $k = 2$.

$$\chi^2_{row_{k=2}} = \frac{\left(13 - \frac{88}{2}\right)^2}{\frac{88}{2}} + \frac{\left(50 - \frac{75}{2}\right)^2}{\frac{75}{2}} + \frac{\left(37 - \frac{37}{2}\right)^2}{\frac{37}{2}} + \frac{\left(75 - \frac{88}{2}\right)^2}{\frac{88}{2}} + \frac{\left(25 - \frac{75}{2}\right)^2}{\frac{75}{2}} + \frac{\left(0 - \frac{37}{2}\right)^2}{\frac{37}{2}} = 89.01 \tag{3}$$

$$\chi^2_{column_{k=2}} = \frac{\left(25 - \frac{205}{3}\right)^2}{\frac{205}{3}} + \frac{\left(80 - \frac{205}{3}\right)^2}{\frac{205}{3}} + \frac{\left(100 - \frac{205}{3}\right)^2}{\frac{205}{3}} + \frac{\left(75 - \frac{95}{3}\right)^2}{\frac{95}{3}} + \frac{\left(20 - \frac{95}{3}\right)^2}{\frac{95}{3}} + \frac{\left(0 - \frac{95}{3}\right)^2}{\frac{95}{3}} = 139.40 \tag{4}$$

Table 5 shows the Chi Index results for our toy example. Chi Index reaches its maximum value at $k = 3$, therefore, we may conclude that the optimal number of clusters that achieved the best clustering solution with this class is with 3 clusters. It should be highlighted that the solution is reached by taking the maximum value of all the solutions because it is the one that achieve the largest value of chi values with both components, and also achieved the minimum difference between them.

Table 4
Toy example contingency tables in which clusters are represented by the rows, and the classes are represented by R (red), G (green), and B(blue). The tables on the left are the contingency tables in absolute values, while tables on the right belongs to the contingency tables with relative values taking as total the sum of the rows (left-side) and the sum of the columns (right-side).

#	R	G	B	
1	2	8	6	16
2	6	2	0	8
	8	10	6	24

(a) $k = 2$.

#	R	G	B	
1	0	2	6	8
2	6	2	0	8
3	2	6	0	8
	8	10	6	24

(b) $k = 3$.

#	R	G	B	
1	0	2	5	7
2	4	1	0	5
3	2	1	0	3
4	2	6	1	9
	8	10	6	24

(c) $k = 4$.

	By row			
#	R	G	B	
1	13%	50%	37%	100%
2	75%	25%	0%	100%
	88%	75%	37%	200%

	By column			
#	R	G	B	
1	25%	80%	100%	205%
2	75%	20%	0%	95%
	100%	100%	100%	300%

(d) Relative contingency tables for $k = 2$.

	By row			
#	R	G	B	
1	0%	25%	75%	100%
2	75%	25%	0%	100%
3	22%	67%	11%	100%
	97%	117%	86%	300%

	By column			
#	R	G	B	
1	0%	20%	100%	120%
2	75%	20%	0%	95%
3	25%	60%	0%	85%
	100%	100%	100%	300%

(e) Relative contingency tables for $k = 3$.

	By row			
#	R	G	B	
1	0%	29%	71%	100%
2	80%	20%	0%	100%
3	67%	33%	0%	100%
4	22%	67%	11%	100%
	169%	149%	82%	400%

	By column			
#	R	G	B	
1	0%	20%	83%	103%
2	50%	10%	0%	60%
3	25%	10%	0%	35%
4	25%	60%	17%	102%
	100%	100%	100%	300%

(f) Relative contingency tables for $k = 4$.

Table 5
Chi index solutions for $k = 2$ to 4.

k	χ^2_{row}	χ^2_{column}	$\chi^2_{row_{\max}}$	$\chi^2_{column_{\max}}$	*Chi Index*(k)
2	89.01	139.40	200	300	0.890
3	277.50	299.38	600	600	**0.925**
4	304.05	237.21	800	600	0.760

3.4. Chi Index definition

The Chi Index is defined as:

$$chi\,index(k) = row_{norm}(k) + col_{norm}(k) - |row_{norm}(k) - col_{norm}(k)| \tag{5}$$

where

$$row_{norm}(k) = \frac{\chi^2_{row}(k)}{\chi^2_{row_{max}}} \tag{6}$$

$$col_{norm}(k) = \frac{\chi^2_{column}(k)}{\chi^2_{column_{max}}} \tag{7}$$

$$\chi^2_{row}(k) = \sum_i^r \sum_j^c \frac{\left(\frac{n_{ij}}{n_{i\cdot}} - \frac{N_{\cdot j}}{r}\right)^2}{\frac{N_{\cdot j}}{r}} \tag{8}$$

$$\chi^2_{column}(k) = \sum_i^r \sum_j^c \frac{\left(\frac{n_{ij}}{n_{\cdot j}} - \frac{N_{i\cdot}}{c}\right)^2}{\frac{N_{i\cdot}}{c}} \tag{9}$$

$$N_{i\cdot} = \sum_j^c \frac{n_{ij}}{n_{\cdot j}} \tag{10}$$

$$N_{\cdot j} = \sum_i^r \frac{n_{ij}}{n_{i\cdot}} \tag{11}$$

and n_{ij} is the number of elements from the cluster i in the class j, $n_{i\cdot}$ is the total number of elements in cluster i, $n_{\cdot j}$ corresponds to the total number of elements in class j, and n is the total of elements in the dataset.

$$\chi^2_{row_{max}} = \begin{cases} 100 \cdot r \cdot (r-1) & r \leq c \\ 100 \cdot r \cdot (c-1) & r > c \end{cases} \tag{12}$$

$$\chi^2_{column_{max}} = \begin{cases} 100 \cdot c \cdot (r-1) & r \leq c \\ 100 \cdot c \cdot (c-1) & r > c \end{cases} \tag{13}$$

where r and c are the number of rows and columns respectively.

Chi index takes a value in [0, 2], where 0 is given by the worst clustering solution, and 2 is the best value that Chi Index can achieve. Hence, the optimal k is given by:

$$k^* = \underset{k}{\operatorname{argmax}}\, chi\, index(k) \tag{14}$$

4. Experimental results

This section describes the experimental study carried out with the aim of testing the proposed Chi Index over a variety of artificial datasets, and 47 public datasets in terms of certain benchmark evaluation criteria.

This section is composed of Section 4.1 that includes the experiments with the synthetic datasets. Section 4.2 defines the experimental design. Section 4.3 presents the results of the experiments carried out with the public datasets. Section 4.3.1 includes a statistical analysis to test the effectiveness of our proposed index for the public datasets. Finally, a discussion of the results is included in Section 4.3.2.

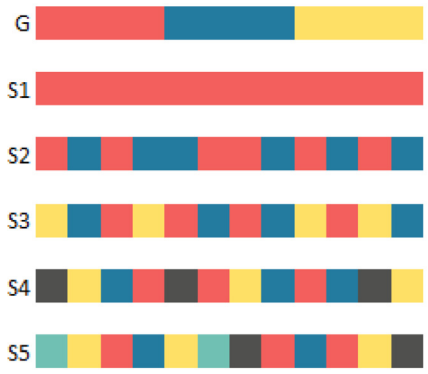
4.1. Chi Index validation

This section includes experimental results for artificial datasets to evaluate the behaviour of Chi Index on diverse clustering solutions based on the work published in [20]. In this case, clustering solutions are generated and compared with the ground truth (G). The results include the 15 CVIs from the state-of-art (Section 2) and our proposed Chi Index. Figs. 5–8 are composed of four subfigures:

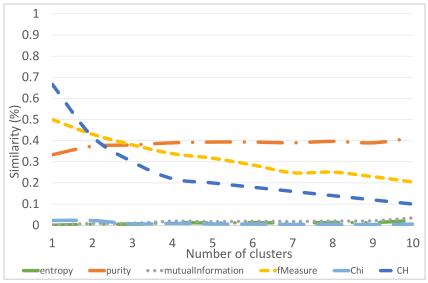
- *Subfigure (a)* is a graphic representation of the generated clustering solutions (S1, S2, S3, ...) with G.
- *Subfigures (b,c,d)* are plots of the CVI results for each of the solutions. The y-axis represents the similarity in percentage, while the x-axis depends on a particular feature of each dataset. Detailed explanations are presented in their respective paragraphs.

The similarity is defined as the affinity measured with the percentage of a clustering solution S_k compared with the ground truth G. It is expressed in relative terms to the best solution that could be found in the interval of the study. Its value lies in the range [0,1], whereby 0 indicates the worst result, and 1 indicates the solution that perfectly fits G.

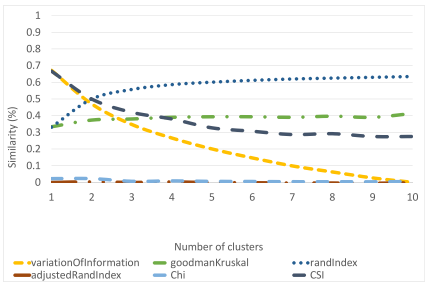
Fig. 5 shows the results for clustering solutions with random partitions. The generated solutions go from 1 class up to 10. Fig. 5a shows the representation of G and the different clustering solutions from 1 class (S1) up to 5 classes (S5). In Figs. 5b–d, it is worth noting that the Chi Index, entropy, mutual information, adjusted rand index, Hubert, and PSI had its



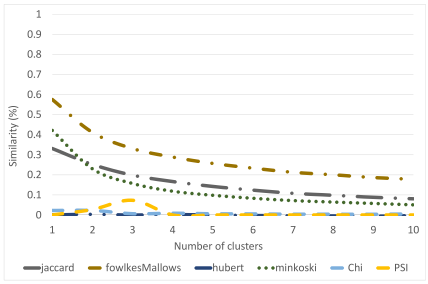
(a) Representation of the ground truth G and the solutions for $k = 1$ to 5 ($S1-S5$).



(b) Solutions for entropy, purity, mutual information, F-measure, CH, and Chi Index.



(c) Solutions for Variation of information, Goodman-Kruskal, Rand Index, Adjusted Rand Index, CSI, and Chi Index.



(d) Solutions for Jaccard, Fowlkes-Mallows, Huber, Minkowski, PSI, and Chi Index.

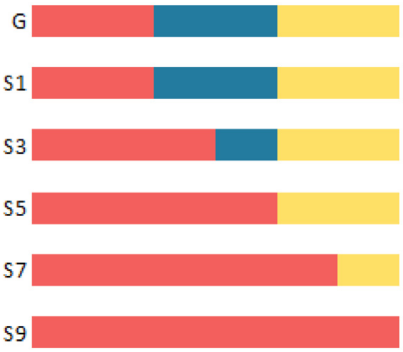
Fig. 5. Results for random generated clustering solution from $k = 1$ to 10 number of clusters.

values at zero. Mutual Information index (Fig. 5d) and the Rand Index (Fig. 5c), could imply that the optimal number is 3 because their curves converge. In addition, PSI had a higher value at 3, that may indicate that is the better solution, but it was with a value under 0.1.

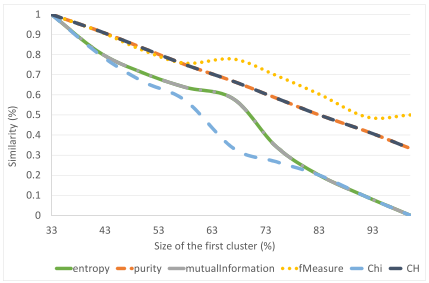
Fig. 6 shows the results for clustering solutions where the instances of the first cluster are increased in each dataset until completion. In Fig. 6a, S1 has the same distribution as G, and hence this is the best solution for all the indices. Figs. 6b–d show the distribution of the CVIs in these datasets. The x-axis represents the percentage of the instances of the first cluster, which ranges from 33% to 100%. It can be observed that all the indices presents a similar behaviour. Their best values are in the dataset that is equal to G and these values decrease until the last dataset whose all instances belong to cluster 1. We find that the Chi Index marks its optimal solution in $S1$ in a similar way than the rest of the indices, but Chi Index descends more linear than the rest of its competitors.

Fig. 7 shows the results for the solutions where the central cluster (in blue) is increased. Fig. 7a shows how the central cluster is increased on each solution where $S1$ is identical to G. The results are similar to the previous ones. Figs. 7b–d show that the indices behave similarly, since the best solution is $S1$, and these indices decrease until the central cluster fills the whole dataset. This result arises from the fact that our index is comparing the distribution of the points across the clusters and, when the dataset is composed of only 1 cluster, the index reaches the lowest value compared with the remaining solutions. We had a comparable situation for the indices of Mutual Information and Entropy (Fig. 7b), Variation of information (Fig. 7c), PSI, and Minkowski (Fig. 7d). It also should be highlighted that Chi Index reached similar results than PSI in this clustering solution.

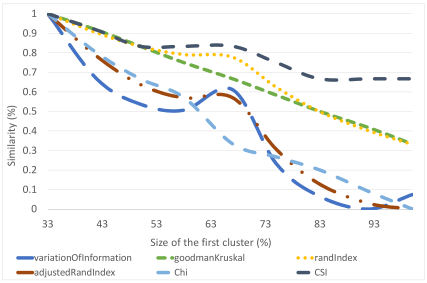
Fig. 8 displays the results of the indices for solutions where the number of incorrect instance labels regularly increases. As seen in Fig. 8a, $S1$ is also identical to G, and it can be observed that on each iteration some of the instances are incorrectly labelled and then this continues until all the instances are incorrectly labelled. Figs. 8b–d show that the Chi Index behaves in a similar way to the rest of the indices during the different datasets. The curves of the indices generally decreases from 1 until 0 in the dataset whose label are 100% incorrect labelled. As it can be seen, the Chi Index and PSI has a near linear since they begin in 1 and decrease to 0. In the case of the F-Measure, the purity, the CSI, and the CH, they start in 1 but they finish at 0.4. The rest of the indices also obtain a similarity of zero in the last dataset but do not describe a near linear behaviour.



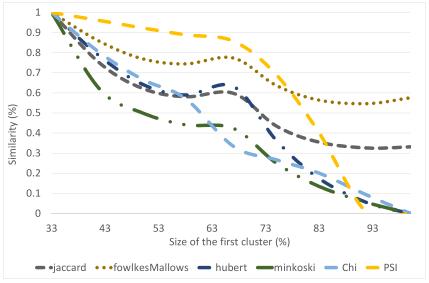
(a) Representation of the ground truth G and the solutions where cluster 1 (in red) is increased until the total of the distribution is reached.



(b) Solutions for entropy, purity, mutual information, F-measure, CH, and Chi Index.



(c) Solutions for Variation of information, Goodman-Kruskal, Rand Index, Adjusted Rand Index, CSI, and Chi Index.



(d) Solutions for Jaccard, Fowlkes-Mallows, Huber, Minkowski, PSI, and Chi Index.

Fig. 6. Results for generated clustering solution where the first cluster increases in each dataset until it fills the whole dataset.

4.2. Experimental design

To generate the clustering solutions, 3 clustering methods from Spark MLlib [19] were applied: k-means, bisecting k-means, and Gaussian mixture.

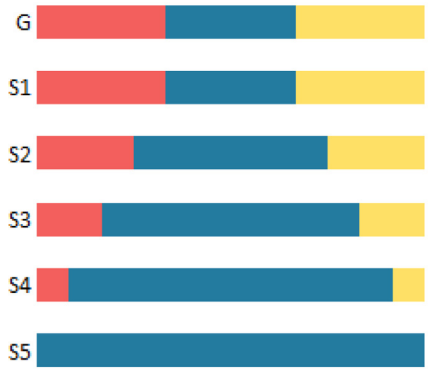
Each dataset, described in Section 4.2.1, was executed with each of these 3 clustering methods. In addition, these clustering methods require the number of clusters (k) into which the dataset is going to be partitioned. The k value was set in the range of $[D_k - 10, D_k + 10]$, where D_k is the correct number of clusters of each dataset and $k > 1$. The number of classes of the datasets was considered as the optimal number of clusters in the same way as carried out in [6,10,20,34,37,39]. With this configuration, we obtained a total of 2820 clustering solutions to test the CVIs. Each clustering solution was compared with the ground truth partition and was then evaluated by the 15 external CVIs described in Section 2. Our new proposed index was also applied in order to compare the results.

4.2.1. Datasets

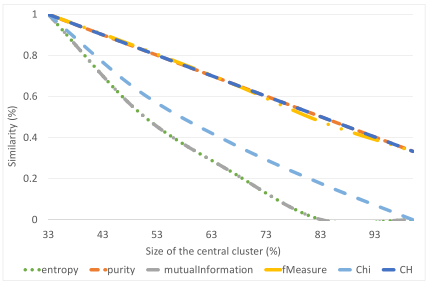
Table 6 presents the datasets used for the experiments and provides the following attributes for each dataset: name; number of classes to be used as the optimal number of clusters; number of features; and the number of instances. All these datasets were downloaded from the UCI machine-learning repository [35]. Note that due to the size of some of the datasets, such as *airlines, higgs, poker*, and *susy*, this could be considered big data. It should be borne in mind that all these datasets included the class information but were not involved in the clustering process. Class information was used in only the clustering analysis stage.

4.2.2. Validity index effectiveness

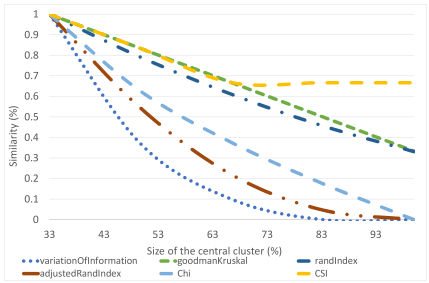
The effectiveness of a CVI measures its capacity to achieve the most coinciding matches while taking a benchmark from different clustering solutions into account. A clustering algorithm and different datasets with a ground truth solution are required in this process. The first step involces applying the clustering algorithm to the datasets and obtaining the multiple solutions. The second step evaluates the solutions with the CVIs. The third step compares the CVI results and selects the one with the highest score.



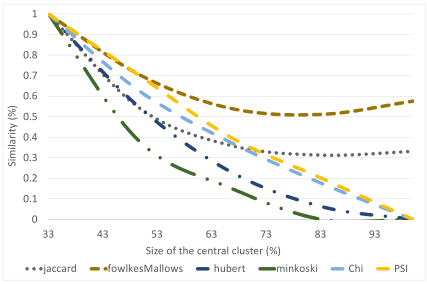
(a) Representation of the ground truth G and the solutions where the central cluster is increased in steps.



(b) Solutions for entropy, purity, mutual information, F-measure, CH, and Chi Index.



(c) Solutions for Variation of information, Goodman-Kruskal, Rand Index, Adjusted Rand Index, CSI, and Chi Index.



(d) Solutions for Jaccard, Fowlkes-Mallows, Huber, Minkowski, PSI, and Chi Index.

Fig. 7. Results for generated clustering solution where the central clusters are increasing step by step until the dataset is completed.

The effectiveness of a CVI depends on how often it takes the correct clustering result in accordance with the chosen criterion. Therefore, the effectiveness is given by counting how many times the index has hit the correct number of clusters. The benchmark employed to make the comparison between the indices includes the following values:

- Average number of hits: this value is given by the mean of the number of times that the index correctly predicted the optimal number of clusters per dataset.
- Average squared error: this is calculated as the average of the squared distances between the prediction of the index I_i and the correct number n_i:

$$Error = \frac{\sum_{i \in n} d(I_i, n_i)^2}{n} \tag{15}$$

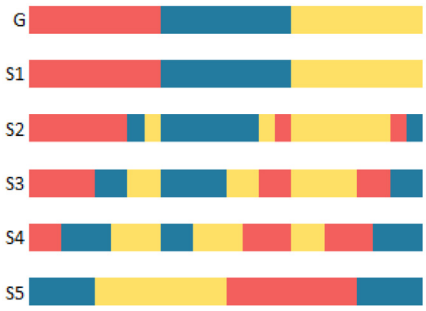
where n is the total number of datasets.

4.2.3. Statistical test

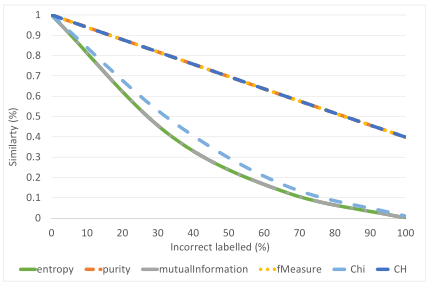
Finally, a statistical framework was applied to test the performance of the indices for the public datasets. The non-parametric Friedman test and Holm post-hoc procedure were chosen to statistically validate the significant differences in the mean ranks of the corresponding p-values reached. This statistical analysis was carried out using the open-source platform StatService [40].

The Friedman test is a non-parametric statistical test that evaluates the differences between more than two related sample means [41]. In our case, the related samples were the CVIs to be compared. The lower the p-value, the better the position in the ranking in the Friedman test.

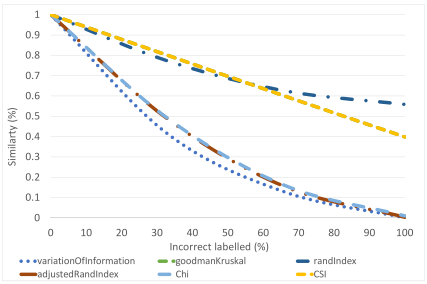
Average ranks for each index provide an objective comparison. The Friedman test could check whether the average ranks were significantly different from the mean rank expected under the null hypothesis. After checking that the measured average ranks are significantly different with an $\alpha = 0.05$, and provided that the Friedman test rejected the null hypothesis, then a post-hoc test could proceed to evaluate the relative performance of the studied CVIs against a control index (that



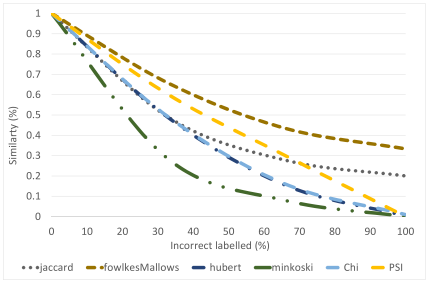
(a) Representation of the ground truth G and the clustering solutions where the number of incorrectly labelled objects are increasing.



(b) Solutions for entropy, purity, mutual information, F-measure, CH, and Chi Index.



(c) Solutions for Variation of information, Goodman-Kruskal, Rand Index, Adjusted Rand Index, CSI, and Chi Index.



(d) Solutions for Jaccard, Fowlkes-Mallows, Huber, Minkowski, PSI, and Chi Index.

Fig. 8. Results for generated clustering solution where the number of incorrectly labelled objects increase proportionally between the clusters.

with the best average rank) thereby avoiding any family-wise errors. This task will be carried out with the Holm step-down procedure by testing hypotheses sequentially ordered in terms of their significance [42].

4.3. Experimental results

This section presents the results obtained with the public datasets. Fig. 9a shows the average number of hits for each CVI in ascending order. It should be highlighted that the Chi Index achieves the highest rate of hits (58%) with a significant margin with its competitors. Indices from the literature had similar rates of hits, ranging from 43% in the case of the F-Measure to 36% for Mutual Information.

On the other hand, Fig. 9b presents the average squared error per index. It is worth noting that the Chi Index obtained the lowest percentage of error. This means that the Chi Index hits the optimal number of clusters most of the times and,when it is in error, it is still not far from the solution.

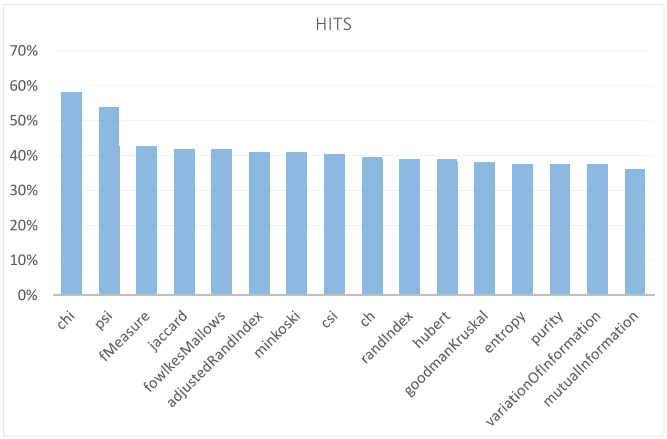
Fig. 10 presents the heatmaps of the distances to the optimal number of clusters of each CVI (rows) for each dataset (columns) represented by the numbers given in Table 6. In these figures, hits are highlighted in green and the farthest results from the solution are graded from white to red. Fig. 10a–c correspond to the results for the k-means, the bisecting k-means and Gaussian mixture methods, respectively.

As can be observed, the Chi Index had a higher rate of green cells than the rest of the CVIs. Although in certain datasets no CVI hit the correct number of clusters, in these cases, the Chi Index remained closer to the solution than its competitors.

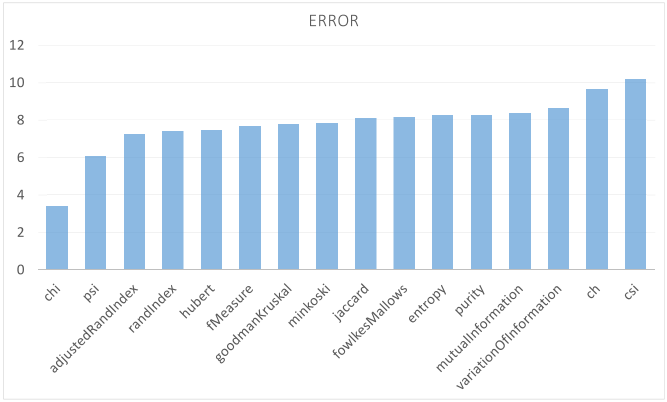
Fig. 11 illustrates the results of Chi Index for two datasets, *zoo* and *knowledge*, whose optimal number of clusters are 7 and 4, respectively. Fig. 11a shows how both curves are crossed at $k = 7$. Moreover, Fig. 11b presents the results for the dataset that has 4 clusters. As can be observed, the curves for the Chi Index by rows and by columns are cut off between $k = 4$ and $k = 5$. These results need no interpretation because the solution is given directly by the index.

4.3.1. Statistical analysis

The Friedman test rankings for every CVI are shown in Table 7a. The ranking was carried out with the results shown in Fig. 10. As previously indicated, the best result for a ranking was 1 and the worst was the last position. As the ranking shows, the Chi Index was in the first position with 6.415. The next index in the ranking was the PSI with a difference of more



(a) Average number of hits by CVI.



(b) Average squared error distance by CVI.

Fig. 9. Benchmark results for the public datasets.

than 1 point with respect to the Chi Index. From this index onwards there are only 0.5 points of difference, and hence, we may conclude that there is a dissimilarity between chi and the indices from the literature. The lowest value for the ranking was 6.415, and the rest ranged from 7.109 to 9.517. Such high values were presented because there were numerous ties in the results, and, in these cases, Friedman establishes the average of the sum of the ranking values of all the competitors. Therefore, for the dataset where all the indices hit the optimal number, Friedman set the ranking at 8.

The statistic for Friedman was 54.694, distributed according to a chi-squared distribution with 15 degrees of freedom. The p-value for Friedman was 0.000, which is lower than 0.05. Therefore, significant differences do exist and it rejected the null hypothesis that they all behaved in a similar way with a level of significance of $\alpha = 0.05$.

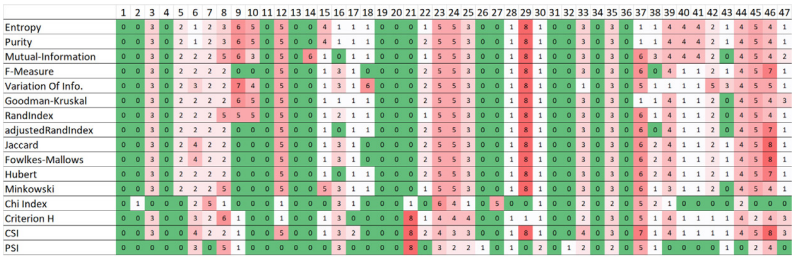
A post-hoc test has been performed in pairs to verify that our proposed Chi Index is significantly different from the other indices.

Table 7b shows the p-values, z-value and α_{Holm}, using the Chi Index as the control method since it obtained the best ranking. As can be observed, the null hypothesis is rejected for all the competitors' CVIs where the $p - value$ remains lower than the α_{Holm}. The null hypothesis was rejected by all the competitors but PSI, whose p-value (0.219) was higher than its α_{Holm} (0.050). Therefore, it may be concluded that the Chi Index generated the best results since it obtained the best average ranking, and that it was significantly different to all the competitors 'CVIs but PSI.

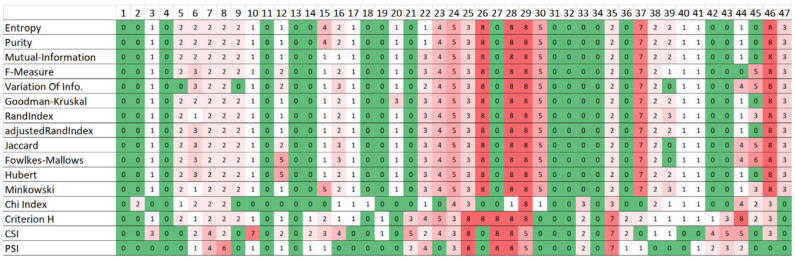
4.3.2. Discussion

The results of the experimental analysis for the public datasets from the UCI repository show that our proposed external index improves the rate of hits by almost 16% (Fig. 9a) with respect to the CVIs from the literature but just 2% from PSI. In addition, in the case of not being able to hit the correct number of clusters, our index obtained a rate of 3 points lower than the CVIs from the literature (Fig. 9b). Chi Index obtained similar rates of hits than PSI, but in case of error, its error is much lower. Our proposed index improves the results based on Friedman's test (Table 7a).

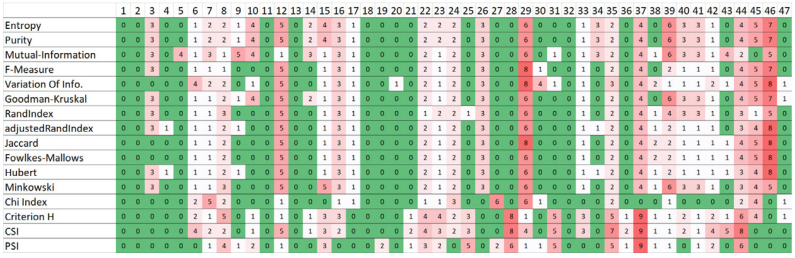
According to the heatmaps from Fig. 10, it can be stated that the Chi Index produced promising results since it hit the optimal number of clusters for most of the datasets and on the according when it failed, its error was not far from the



(a) Heatmap for k-means results.

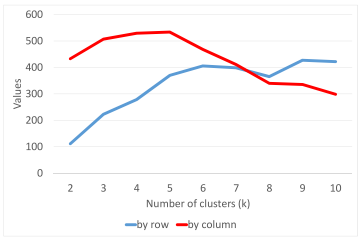


(b) Heatmap for bisecting k-means results.

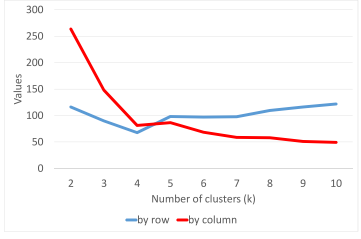


(c) Heatmap for Gaussian mixture results.

Fig. 10. Heatmaps of the distances to the optimal number of clusters of each CVI (rows) for each dataset (columns) represented by the number given in Table 6.



(a) Solution for zoo dataset whose optimal number of clusters is 7.



(b) Solution for knowledge dataset whose optimal number of clusters is 4.

Fig. 11. Representation of the Chi Index for $k = 2$ to 10 for two real datasets.

optimal. It is also interesting to note that there were several datasets in which none of the indices hit the optimal number of clusters. However, in numerous of datasets, it was only the Chi Index which hit the optimal number of clusters.

If we analyse the rate of hits and errors per clustering method, then the Chi Index obtained the best values. For k-means, the Chi Index and the PSI attained 60% hits, and 3.49 and 3.98 points of error respectively. The third in the ranking was the CH index with 49% hits and 5.68 points of error. Bisecting k-means results show that the Chi Index had the highest rate of hits with 64%, while the second mark was obtained by several indices with 49%. The Chi Index had 3.11 points of error and the next in the ranking was the Rand index with 4.32. Finally, the Gaussian mixture had similar results. The PSI index had 53% hits and 8.00 points of error, and in the second position was the Chi Index with 51% hits and 3.53 points of error. K-means and bisecting k-means obtained similar results while Gaussian mixture solutions obtained a lower rate of hits and a higher error.

Table 6
Dataset description.

#	Dataset	Classes	Features	Instances
1	airlines	2	7	539,383
2	bankmarketing	2	16	45,228
3	banknote	2	4	1372
4	biodeg	2	41	1055
5	breast cancer wisconsin	2	9	699
6	breast-tissue	6	9	106
7	car	4	6	1728
8	cloud	4	10	1024
9	column_2C	2	6	310
10	column_3C	3	6	310
11	diabetes	2	20	768
12	ecoli	8	7	336
13	electricity	2	8	45,312
14	faults	2	27	1941
15	forest type mapping	4	27	523
16	gesture phase dataset	5	32	9873
17	glass	6	9	214
18	haberman	2	3	306
19	higgs	2	28	11,000,000
20	iris	3	4	150
21	kddcup99	2	41	494,020
22	knowledge	4	5	403
23	leaf	36	14	340
24	letter	26	16	20,000
25	movement	15	90	360
26	optdigits	10	64	5620
27	ozone	2	72	2534
28	pendigits	10	16	10,992
29	poker	10	10	829,202
30	relax	2	13	182
31	satimage	7	36	6435
32	seeds	3	7	210
33	segment	7	19	2310
34	spambase	2	57	4601
35	spectrometer	4	100	531
36	susy	2	12	5,000,000
37	urban land cover	9	147	675
38	vehicle	4	18	846
39	vowel	11	10	990
40	waveform-1	3	21	5000
41	waveform-2	3	40	5000
42	wholesale	2	7	440
43	wine	3	13	178
44	wine quality red	6	11	1599
45	wine quality white	7	11	4898
46	yeast	10	8	1484
47	zoo	7	17	101

It is also interesting to note that the Chi Index illustrates the optimal clustering solution in an easy and concise way. Some of the solutions of indices in the literature need to be interpreted by following the elbow method or looking for a minimum or a maximum. The Chi Index points out the optimal solution in the intersection of the described curves.

5. Conclusions

In this paper, an innovative external CVI implemented in Spark has been proposed for its application in datasets regardless of their size. The proposed Chi Index is based on the chi-squared statistic test. In addition, we have shown the differences between our proposal and the indices from the literature.

The experimental study indicates that our external index is very competitive. Its effectiveness in public datasets with different sizes has been tested while varying the number of clusters, features, and the number of instances. The main achievements include the following:

- An external CVI based on the chi-squared statistic test is given.
- Our index allowed us to estimate the optimal number of clusters based on the class of the dataset.
- Chi-index results are clear to read and require no further interpretation.
- The proposed index is equipped to work with datasets that may be considered as Big Data.

Table 7
Statistical results.

(a) Sorted mean ranking for Friedman's test.		(b) Post-hoc analysis using Holm procedure and the Chi Index as the control index.			
CVI	Ranking	CVI	p	z	α_{Holm}
Chi Index	6.415	CSI	0.0000	5.490	0.0033
PSI	7.109	Variation of Information	0.0000	4.792	0.0036
CH	8.151	Purity	0.0000	4.543	0.0038
Adjusted Rand Index	8.383	Mutual Information	0.0000	4.493	0.0042
F-Measure	8.415	Entropy	0.0000	4.462	0.0045
Rand Index	8.489	Jaccard	0.0000	4.219	0.0050
Minkowski	8.545				
Hubert	8.640	Fowlkes–Mallows	0.0000	4.187	0.0056
Goodman-Kruskal	8.753	Goodman–Kruskal	0.0000	4.137	0.0063
Fowlkes-Mallows	8.781	Hubert	0.0001	3.938	0.0071
Jaccard	8.799	Minkowski	0.0002	3.770	0.083
Entropy	8.936	Rand Index	0.0002	3.670	0.0100
Mutual Information	8.954	F-Measure	0.0004	3.539	0.0125
Purity	8.982	Adjusted Rand Index	0.0005	3.486	0.0167
Variation of Information	9.123	CH	0.0021	3.486	0.0250
CSI	9.517	PSI	0.2197	1.227	0.0500

- The size of the dataset does not directly influence the effectiveness of the index.
- The software of this contribution can be found as a spark-package at http://spark-packages.org/package/josemarialuna/ExternalValidity.
- The source code of the Chi Index and the other indices from the literature can be found at https://github.com/josemarialuna/ExternalValidity.

We are currently applying this Chi Index in a clustering analysis with employment data and promising results have been attained. The Chi Index is also being applied on electrical data in collaboration with a Spanish electricity company. As future work, it would be interesting to extend the application of the index to include multi-label datasets.

Acknowledgment

This work has been supported by the Spanish Ministry of Economy and Competitiveness under projects TIN2014-55894-C2-R and TIN2017-88209-C2-2-R. J.M. Luna-Romera holds a FPI scholarship from the Spanish Ministry of Economy and Competitiveness.

References

[1] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (8) (2010) 651–666.
[2] M. Castro-Franco, M. Córdoba, M. Balzarini, J. Costa, A pedometric technique to delimitate soil-specific zones at field scale, Geoderma 322 (2018) 101–111.
[3] R. Davoodi, M. Moradi, Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier, J. Biomed. Inform. 79 (2018) 48–59.
[4] R. Pérez-Chacón, J.M. Luna-Romera, A. Troncoso, F. Martínez-Álvarez, J.C. Riquelme, Big data analytics for discovering electricity consumption patterns in smart cities, Energies 11 (3) (2018).
[5] B. Zhao, J. Wang, Unification of particle velocity distribution functions in gas-solid flow, Chem. Eng. Sci. 177 (2018) 333–339.
[6] J. Rojas-Thomas, M. Santos, M. Mora, New internal index for clustering validation based on graphs, Expert Syst. Appl. 86 (2017) 334–349.
[7] J. Handl, J. Knowles, D.B. Kell, Computational cluster validation in post-genomic data analysis, Bioinformatics 21 (15) (2005) 3201–3212.
[8] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2) (2001) 107–145.
[9] J. Wu, H. Xiong, J. Chen, Adapting the right measures for K-means clustering, in: Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD, ACM, New York, NY, USA, 2009, pp. 877–886.
[10] C. Liu, W. Wang, M. Konan, S. Wang, L. Huang, Y. Tang, X. Zhang, A new validity index of feature subset for evaluating the dimensionality reduction algorithms, Knowl. Based Syst. 121 (2017) 83–98.
[11] S. Jabbar, A.A. Minhas, A. Paul, S. Rho, Multilayer cluster designing algorithm for lifetime improvement of wireless sensor networks, J. Supercomput. 70 (1) (2014) 104–132.
[12] R. Tibshirani, G. Walther, Cluster validation by prediction strength, J. Comput. Graph. Stat. 14 (3) (2005) 511–528.
[13] A. Paul, A. Ahmad, M.M. Rathore, S. Jabbar, Smartbuddy: defining human behaviors using big data analytics in social internet of things, IEEE Wirel. Commun. 23 (5) (2016) 68–74.
[14] V. Berikov, I. Pestunov, Ensemble clustering based on weighted co-association matrices: error bound and convergence properties, Pattern Recognit. 63 (2017) 427–436.
[15] Y. Lei, J.C. Bezdek, S. Romano, N.X. Vinh, J. Chan, J. Bailey, Ground truth bias in external cluster validity indices, Pattern Recognit. 65 (2017) 58–70.
[16] E. López-Rubio, E.J. Palomo, F. Ortega-Zamorano, Unsupervised learning by cluster quality optimization, Inf. Sci. 436–437 (2018) 31–55.
[17] H. Yahyaoui, H.S. Own, Unsupervised clustering of service performance behaviors, Inf. Sci. 422 (2018) 558–571.
[18] Y. Zhang, J. Madziuk, C.H. Quek, B.W. Goh, Curvature-based method for determining the number of clusters, Inf. Sci. 415–416 (2017) 414–428.
[19] A. Spark, Clustering - Spark 2.2.0 Documentation, 2018. https://spark.apache.org/docs/2.2.0/ml-clustering.html, [Online; accessed 6-april-2018].
[20] M. Rezaei, P. Fränti, Set matching measures for external cluster validity, IEEE Trans. Knowl. Data Eng. 28 (8) (2016) 2173–2186.
[21] Y. Zhao, G. Karypis, Criterion functions for document clustering: experiments and analysis, Technical Report, University of Minnesota, Department of Computer Science, Minneapolis, 2001.
[22] B. Larsen, C. Aone, Fast and effective text mining using linear-time document clustering, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD, ACM, New York, NY, USA, 1999, pp. 16–22.

[23] M. Meilă, D. Heckerman, An experimental comparison of model-based clustering methods, Mach. Learn. 42 (1) (2001) 9–29.
[24] P. Fränti, M. Rezaei, Q. Zhao, Centroid index: cluster level similarity measure, Pattern Recognit. 47 (9) (2014) 3034–3045.
[25] L.A. Goodman, W.H. Kruskal, Measures of Association for Cross Classifications, Springer New York, New York, NY, 1971, pp. 2–34.
[26] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (336) (1971) 846–850.
[27] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary? in: Proceedings of the Twenty Sixth Annual International Conference on Machine Learning, in: ICML, ACM, New York, NY, USA, 2009, pp. 1073–1080.
[28] R. Sokal, P. Sneath, Principles of Numerical Taxonomy, Books in biology, W. H. Freeman, 1963.
[29] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, J. Am. Stat. Assoc. 78 (383) (1983) 553–569.
[30] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985) 193–218.
[31] A. Ben-Hur, I. Guyon, Detecting stable clusters using principal component analysis, in: M.J. Brownstein, A.B. Khodursky (Eds.), Functional Genomics: Methods and Protocols, Humana Press, Totowa, NJ, 2003, pp. 159–182.
[32] M. Meilă, Comparing clusterings by the variation of information, in: B. Schölkopf, M.K. Warmuth (Eds.), Learning Theory and Kernel Machines, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 173–187.
[33] A. Banerjee, I.S. Dhillon, J. Ghosh, S. Sra, Clustering on the unit hypersphere using von Mises-Fisher distributions, J. Mach. Learn. Res. 6 (2005) 1345–1382.
[34] D. Campo, G. Stegmayer, D. Milone, A new index for clustering validation with overlapped clusters, Expert Syst. Appl. 64 (2016) 549–556.
[35] D. Dheeru, E. Karra Taniskidou, UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml/citation_policy.html.
[36] A.K. Alok, S. Saha, A. Ekbal, A min-max distance based external cluster validity index: MMI, in: Proceedings of the Twelfth International Conference on Hybrid Intelligent Systems (HIS), 2012, pp. 354–359.
[37] J. Rodríguez, M. Medina-Pérez, A. Gutierrez-Rodríguez, R. Monroy, H. Terashima-Marín, Cluster validation using an ensemble of supervised classifiers, Knowl. Based Syst. 145 (2018) 1–14.
[38] P.E. Greenwood, M.S. Nikulin, A guide to chi-squared testing, Wiley-Interscience, New York, NY, 1996.
[39] M.A. Wani, R. Riyaz, A new cluster validity index using maximum cluster spread based compactness measure, Int. J. Intell. Comput. Cybern. 9 (2) (2016) 179–204.
[40] J.A. Parejo, J. Garcia, A. Ruiz-Cortes, J.C. Riquelme, Statservice: Herramienta de análisis estadístico como soporte para la investigación con metaheurísticas, Actas del VIII Congreso Expañol sobre Metaheurísticas, Algoritmos Evolutivos y Bio-inspirados, 2012.
[41] S. Garcia, J. Luengo, F. Herrera, Data Preprocessing in Data Mining, Springer Publishing Company, Incorporated, 2014.
[42] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (2) (1979) 65–70.

Capítulo 6

Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities

Resumen

En este artículo se presenta una novedosa metodología para analizar el consumo eléctrico de una smart city aplicando técnicas de minería de datos con tecnología Big Data. Esta metodología tiene como objetivo ayudar en la fase de toma de decisión para ahorrar costes económicos y en energía. La metodología propuesta se compone de 4 fases: en la primera fase se preparan los datos obtenidos mediante sensores y se preprocesan para su posterior análisis; en la segunda fase se calcula el número óptimo de *clusters* del conjunto de datos aplicando cuatro índices de validación de clustering, cuyos resultados se valorarán teniendo en cuenta un sistema de votación; el algoritmo de clustering, en este caso, k-means, se aplica en la tercera fase, quedando agrupados aquellos datos que compartan mayor similitud; y por último, estos resultados se analizan y se caracterizan los *clusters* ofreciendo una visión específica de los datos que componen esos *clusters*. Los resultados de aplicar esta metodología se muestran en tablas y gráficas de fácil interpretación y permitirá al usuario final tomar decisiones en base a ellos. La experimentación ha sido llevada a cabo usando datos del consumo eléctrico de 8 edificios la Universidad Pablo de Olavide entre los años 2011 y 2017. Además, se ha calculado el rendimiento de esta metodología aplicándola a datos sintéticos para simular una smart city compuesta de 120.000 edificios.

- Estado: Publicado en Energies (MDPI) (2018), Volumen: 11(3), 683

- Índice de Impacto (JCR 2018): 2.707
- Área de Conocimiento:
  - Energy and Fuel. Ranking 56/103 - Q3
- Citas:
  - Scopus: 17
  - Google Scholar: 22
  - Web of Science: 13

energies

MDPI

Article

Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities

Rubén Pérez-Chacón [1,†], José M. Luna-Romera [2,†], Alicia Troncoso [1,*], Francisco Martínez-Álvarez [1] and José C. Riquelme [2]

[1] Division of Computer Science, Universidad Pablo de Olavide, ES-41013 Seville, Spain; rpercha@upo.es (R.P.-C.); fmaralv@upo.es (F.M.-Á.)
[2] Division of Computer Science, University of Sevilla, ES-41012 Seville, Spain; jmluna@us.es (J.M.L.-R.); riquelme@us.es (J.C.R.)
* Correspondence: atrolor@upo.es; Tel.: +34-954-349-230
† These authors contributed equally to this work.

Received: 31 January 2018; Accepted: 13 March 2018; Published: 18 March 2018

Abstract: New technologies such as sensor networks have been incorporated into the management of buildings for organizations and cities. Sensor networks have led to an exponential increase in the volume of data available in recent years, which can be used to extract consumption patterns for the purposes of energy and monetary savings. For this reason, new approaches and strategies are needed to analyze information in big data environments. This paper proposes a methodology to extract electric energy consumption patterns in big data time series, so that very valuable conclusions can be made for managers and governments. The methodology is based on the study of four clustering validity indices in their parallelized versions along with the application of a clustering technique. In particular, this work uses a voting system to choose an optimal number of clusters from the results of the indices, as well as the application of the distributed version of the k-means algorithm included in Apache Spark's Machine Learning Library. The results, using electricity consumption for the years 2011–2017 for eight buildings of a public university, are presented and discussed. In addition, the performance of the proposed methodology is evaluated using synthetic big data, which cab represent thousands of buildings in a smart city. Finally, policies derived from the patterns discovered are proposed to optimize energy usage across the university campus.

Keywords: big data; time series clustering; patterns; smart cities

1. Introduction

Governments in many metropolises are embracing the concept of smart cities, and are beginning to collect big datasets in order to obtain valuable information from them. This information helps governments to improve the standards of living and sustainability required for their inhabitants. In order to increase the comfort and life quality of citizens, it is necessary to reduce costs and optimize the consumption of different energy resources. This reduction in costs, for instance, could improve performance in areas such as education, health-care, transport, security, and emergency services [1]. In this regard, massive storage of data using smart grid technologies is widespread [2]. For example, the energy consumption of water or electricity in public institutions is continuously monitored. However, traditional tools and techniques for storing and extracting valuable information have become obsolete due to the high computational cost of mining gigabytes of data [3]. In this sense, the advent of new machine learning tools makes it easier to mine data, but new techniques are needed to improve the processing, management, and discovery of valuable information and knowledge for organizations [4].

Given the sudden need to process and extract valuable information for organizations, the MapReduce paradigm [5] emerged in the context of distributed computing applications. Later, an open source paradigm called Apache Spark [6] appeared, with the fault tolerance of MapReduce but more significant capabilities such as multi-step computing or the use of high-level operators and various programming languages. It is worth mentioning the optimization of this technology using the Scala language and the Resilient Distributed Dataset (RDD) variables [7], as well as the integration of the Machine Learning Library (MLlib) in the framework [8].

The aim of this work is the active treatment and discovery of electricity consumption patterns from big data time series. Due to the large size of the datasets, modern machine learning techniques based on distributed computing will be used to analyze the data. In this sense, we propose a methodology that optimizes the use of the parallelized version of k-means [9] by studying several cluster validation indices (CVIs) [10], some of which are computationally designed to process big data [11]. A vote-based strategy using the variety of outcomes obtained by these CVIs is proposed [12].

This work draws valuable conclusions from the analysis and study of the consumption patterns of a big data time series of electricity consumption of several buildings of Pablo de Olavide University, extracted using smart meters over six years. Besides, the size of the initial dataset has been multiplied in such a way so as to demonstrate the usefulness and efficiency of the methodology proposed for use in the context of smart cities. It is expected that this methodology will be used to characterize electricity consumption over time and results will be useful for making decisions regarding the efficient use of energy resources.

The rest of the paper is structured as follows. Section 2 describes the related work, and Section 3 proposes the methodology used to uncover patterns in big data time series. Section 4 presents the experimental results for data pre-processing, the study of CVI, and the application of the parallelized k-means algorithm. Finally, Section 6 summarizes the main findings of the study.

2. Related Work

Electricity consumption has soared in recent years to levels never before seen, as cities and countries have advanced technologically. If this demand for energy is no longer met by individual governments at the global level, the problems caused by climate change may increase.

In the last years, many works have been published on this issue in the context of smart cities. A review of the development of smart grid technologies with a view to energy conservation and sustainability can be found in [13]. A smart city can be defined as an efficient and sustainable urban centre that assures high quality of life by optimizing its resources. Energy management is one of the most demanding issues within these urban centres. A methodology to develop an improved energy model in the context of smart cities is proposed in [14]. The concept of smart communities is defined in [15] as the union of several cities that implement and take advantage of these technologies, with the objective of improving the habitability, preservation, revitalization, and affordability of a community. Attention has also been recently paid to the optimization of electrical networks through the installation of smart meters, used for data collection in this work. A study on the unification of smart grids with an energy cooperation approach can be found in [16].

Multiple studies to determine electrical profiles for small and medium-sized assemblies using clustering techniques have been published in the literature. The authors of [17] propose obtaining clusters using a visualization-based methodology. Patterns associated with seasons and days of the year with respect to electricity prices in the Spanish market were discovered in [18]. This article proposed the application of crisp clustering techniques, contrasting the fuzzy clustering methodology evaluated in [19]. In [20] the information provided by clustering techniques was used as input parameters for forecasting consumption. Electrical data from industrial parks were used to apply classification and grouping of patterns in [21]. This work was based on the application of the k-means algorithm and the cascading application of self-organized maps to introduce a computer system that predicted energy consumption patterns in Spanish industrial parks.

However, clustering techniques applied to large quantities of data have taken on importance in recent years. A survey on this subject can be found in [22]. Specifically, several approaches to clustering big data time series have been recently proposed. In [23], the authors suggested a new clustering algorithm based on a previous clustering of a sample of the input data. The similarity among large series was tested by studying the dynamic deformation of time in [24]. A parallel version of k-means using MapReduce technology was applied to obtain clusters of medium-sized datasets in [25]. A distributed method for the initialization of k-means was proposed in [26], but very few works have been published in this regard. The Gaussian mixed model was used to apply clustering to a dataset extracted from smart meters installed in Irish households for a year, studying socio-economic relations and making conclusions based on consumption behaviours [27].

On the other hand, the forecasting of the energy consumption of buildings and campuses has an immense value for energy efficiency and sustainability in the context of smart cities. An important and recent survey [28] thoroughly reviewed the existing machine learning techniques for forecasting the energy consumption of time series. The authors of [29] proposed data clustering and frequent pattern analysis on energy time series to predict energy usage, achieving an acceptable accuracy. Building energy consumption prediction was also applied in [30]. In particular, deep learning techniques, such as autoencoders, were applied to a dataset composed of 8734 instances, reporting great results. Most of these forecasting techniques use the results obtained by a clustering technique as a previous step. However, none of the clustering methods used for the prediction algorithms were used in a parallel and distributed way using a very large set of input data, to the best of our knowledge. Therefore, this work intends to provide a reliable, fast, and accurate clustering method as the basis for these forecasting algorithms dealing with big data time series, and in addition develop a methodology to detect patterns of energy consumption from big data time series collected by sensors in buildings of a smart city.

3. Methodology

This section describes the methodology proposed with the aim of finding patterns of electricity consumption in big data time series. In particular, this methodology obtains electricity consumption patterns by studying the resulting clusters provided by the k-means included in the Machine Learning Library of Apache Spark.

The key steps of the proposed methodology for obtaining consumption patterns are shown in Figure 1.

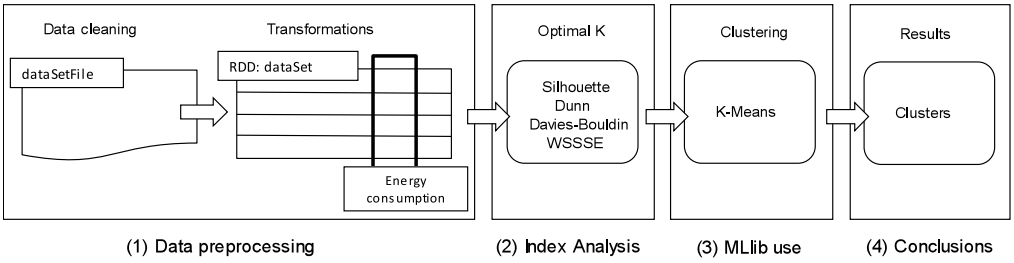


Figure 1. Proposed methodology. RDD: Resilient Distributed Dataset; MLlib: Machine Learning Library; WSSSE: Within Set Sum of Square Errors.

3.1. First Phase: Data Preprocessing

The first phase consists in data preprocessing. The objective of this phase is to clean and perform transformations in the original dataset to create a RDD variable, which can be distributed in a cluster and processed by Spark. The original dataset was obtained from the processing of several CSV files. These files contained records in the form of time series of power consumption data from six buildings of a public university. Data were extracted from the smart meters installed in the buildings. The smart

meters collected electricity consumption records every 15 min from 2011 to 2016. Each row of the starting RDD variable is composed of five values: the name of the building, the date and time (separated into five values), and the energy consumption data at that time. In the data cleansing phase, our application pre-processed the rows containing missing records and accumulated consumption data so that correct learning models could be created in the next phases. This cleaning phase will be discussed extensively in the results section.

Before creating the model, it is necessary to perform a transformation in the original dataset by grouping the energy consumption series into rows of 96 records corresponding to a day. As each hour of the day contains four measurements, the original dataset has a total of 823,776 records, which are grouped per day generating a set with a total of 8581 instances.

In order to be able to identify which day a given result belongs to after applying clustering techniques, we will enter a unique identifier for each instance. This identifier is defined by combining the name of the building with the numerical date on which the measurements were taken.

Thus, each row of the RDD will finally contain a unique identifier and the 96 electric consumption records, in order to obtain conclusions associated with a particular day and building.

3.2. Second Phase: Obtaining the Optimal Number of Clusters

The second phase of the methodology consists in obtaining the optimal number of clusters for the dataset by analysing and interpreting various CVIs. However, some CVIs have limitations to be applied to large datasets due to the computational costs of quadratic complexity. This cost could take much longer to apply than the clustering algorithm used in this study. For this reason, we have applied big data clustering validity indices (BD-CVIs) [11].

In this paper we analyze the results of four BD-CVIs. Three of them are based on traditional CVIs—the BD-Silhouette, BD-Dunn and Davies-Bouldin indices—and the other is based on the Within Set Sum of Square Error (WSSSE) index offered by the MLlib. These BD-CVIs will be defined below. Let Ω be the space of the objects with a given distance d. Let $\{A_k\}_{k=1..N}$ be a set of clusters so that $\bigcup_k A_k = \Omega$ and $A_i \cap A_j = \emptyset \quad \forall i \neq j$. Let C_k be the centroid of A_k and C_0 the centroid of Ω.

BD-Silhouette: This index [11] is defined as the difference between inter-cluster and intra-cluster distances, divided by the maximum of them. The inter-cluster distance is the average of distances between each cluster centroid and global centroid C_0. It is defined by:

$$inter\text{-}cluster = \frac{1}{N}\sum_{k=1}^{N} d(C_k, C_0) \tag{1}$$

The intra-cluster distance is defined as the average of the sum of the distances between each point and the centroid of the cluster to which it belongs. It is defined by the following equations:

$$r_k = \frac{1}{|A_k|}\sum_{x_i \in A_k} d(x_i, C_k) \tag{2}$$

$$intra\text{-}cluster = \frac{1}{|N|}\sum_{k=1}^{N} r_k \tag{3}$$

Therefore, the BD-Silhouette index is defined as follows:

$$BD\text{-}Silhouette = \frac{inter\text{-}cluster - intra\text{-}cluster}{max\{inter\text{-}cluster, intra\text{-}cluster\}} \tag{4}$$

This value can range from -1 to 1 depending on the separation and consistency of the clusters. At the negative end, it will take the value -1 when there is only one cluster, and at the positive end, it will take the value 1 when there is a cluster for each of the dataset elements. In order to find an

optimal value, it is necessary to look for the lowest possible K that maximizes the coherence and consistency of the cluster, being this the first maximum of the BD-Silhouette index.

BD-Dunn: This index [11] relates the maximum distance between all the points belonging to the same cluster and its corresponding centroid, and the minimum distance between these centroids and the global centroid.

$$BD\text{-}Dunn = \frac{\min\limits_{k=1..N}\{d(C_k, C_0)\}}{\max\limits_{k=1..N}\max\limits_{x_i \in A_k}\{d(x_i, C_k)\}} \tag{5}$$

This value is 0 if there is only one cluster and tends to zero when the number of clusters increases. Therefore, a maximum value in the BD-Dunn graph implies a higher quality of the clusters.

Davies-Bouldin: This index [31] assesses how distant clusters can be in order to make them of higher quality. Therefore, we will choose the first minimum of the Davies-Bouldin value chart to create a better model. The index is defined as follows:

$$Davies\text{-}Bouldin = \frac{1}{N}\sum_{i}^{N}\sum_{j}^{N}\max_{i \neq j}\frac{r_i + r_j}{d(C_i, C_j)} \tag{6}$$

where r_i and r_j are represented in Equation (2), and $d(C_i, C_j)$ is the distance between the centroids C_i and C_j.

Within Set Sum of Square Errors (WSSSE): This index [32] is implemented in the MLlib. It is a measure of cluster cohesiveness and it calculates the sum of the distances from each point to the centroid of its cluster.

$$WSSSE = \sum_{x_i \in A_k} d(x_i, C_k)^2 \tag{7}$$

The optimal k is generally the one with a global minimum or the result after applying the "elbow method" to the WSSSE graph [33].

Majority Voting Methodology

The aim is to apply this group of indices to the complete set of data so that we can validate them and also obtain the optimal number of clusters K, which will be used as an input parameter of the parallelized k-means algorithm. In this sense, this work proposes a methodology of majority voting [34], which combines the results obtained of the application of the four indices above as a single result.

The voting strategy is now explained. The application of each of the indices separately to the complete dataset generates a graph that will show maxima or minima indicating the optimal number of clusters according to each case. Therefore, each of the graphs will have a first best value, a second best value, a third best value, and so on.

The voting system will evaluate the best results of all indices so that we extract as a result of the optimal k number of clusters to group our dataset.

There is a favourable and ideal situation: that all indices coincide with the best value or that most (i.e., at least three) coincide. In this case, we will take this value as the optimum k.

However, a second situation may arise: there are no coincidences or these are the minority (that is, fewer than three). When this case occurs, in addition to the first best values, the second best results of the four indices will be also considered. If a majority is not reached, we will study the third best results, and so on until we find a majority that matches. The selected k will then be the one that is repeated most times until the majority is found.

An example of the application of this system is shown in Table 1. In this case, the second situation occurs: only the BD-Dunn (six clusters) and Davies-Bouldin (six clusters) indices offer the same result (i.e., the minority of the first best results), resulting in the application of BD-Silhouette (four clusters) and WSSSE (seven clusters) indices in a manner different from the first two. At this point, we will have to look at the second best results to obtain the optimal number of clusters. If we look at the second best values, BD-Silhouette index coincides with BD-Dunn and Davies-Bouldin indices, since it offers six clusters as the second best result. Therefore, we will have found a majority (i.e., at least three matches), observing the first and second best results of the validity indices.

Table 1. Majority voting methodology.

Values	BD-Silhouette	BD-Dunn	Davies-Bouldin	WSSSE
First	4	**6**	**6**	7
Second	**6**	8	9	15
Third	9	13	15	21

3.3. Third Fase: MLlib

Once the optimal number of clusters k for the dataset has been obtained, the clustering algorithm can be applied. The algorithm used for discovering patterns from the dataset is the k-means [9]. This algorithm is a parallelized version of the k-means included in the MLlib of Apache Spark. This clustering algorithm is based on the classic k-means algorithm and has been developed to extract patterns in parallel and distributed systems.

Figure 2 shows how a run of the k-means works. First, the RDD object containing the complete dataset is distributed in several slave nodes for the execution of k-means, obtaining initial centroids n. Second, the Apache Spark engine shuffles the resulting n centroids for each run. Finally, the k-means algorithm computes the WSSSE index in each partition for each centroid, returning the one that minimizes the WSSSE as the best. It is worth remembering that there are as many centroids as there are concurrent executions. Figure 2 is representative of one execution.

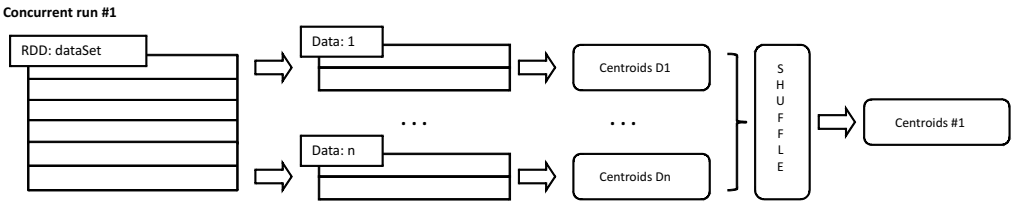


Figure 2. One concurrent execution of the k-means algorithm.

3.4. Fourth Phase: Evaluation

The last phase corresponds to interpret and evaluate the results obtained after the application of k-means with the optimal number of clusters to the dataset.

We will obtain and analyze different types of results to obtain electric consumption patterns in big data time series. We will obtain the distribution of instances in each cluster and the centroids of the daily electricity consumption clusters.

Although clustering is considered an unsupervised learning technique, a clustering validity analysis has been carried out in this study, using features of the instances such as a type of day, season, or building as labels. Clustering results are merged with the features that each instance could have. Table 2 shows an example of the data that will be analysed Each row represents an instance of the dataset, i.e., the electricity consumption of a day, and the column cluster indicates the cluster assigned to that consumption. Also, each instance includes the features to be analysed as the building in which electricity was consumed, the season of the year, and the day of the week or non-working day.

Table 2. Example of a dataset along with assigned cluster and features.

ID	Cluster	Building	Season	Day
1	1	Build_1	Summer	Day off
2	1	Build_1	Winter	Day off
3	2	Build_20	Summer	Thursday
4	1	Build_42	Summer	Friday
5	3	Build_1	Autumn	Monday

With this information, we can see how the clusters are built regarding the features. Following our example, we can observe how the buildings are distributed by clusters, check out in which cluster there are more days off, or determine if the clusters are influenced by the season of the year. Based on this reasoning, we will draw the general conclusions using percentages of the distribution of buildings in the k clusters.

We will also conduct a study of several synthetic big datasets. Starting from the base of the original set, we will multiply its original size with the objective of checking the efficiency in computing time of the proposed methodology.

All the experiments were executed in Amazon Web Services (AWS) Elastic Map Reduce using two different hardware scenarios:

- Five instances of *m*3.*xlarge* with Intel Xeon E5-2670 v2 (Ivy Bridge) processors with 8 CPUs, 15 GB RAM, and 2 SSDs of 40 GB each.
- Five instances of *m*3.2*xlarge* with Intel Xeon E5-2670 v2 (Ivy Bridge) processors with 16 CPUs, 30 GB RAM, and 2 SSDs of 80 GB each.

4. Results

This section is organized as follows. Section 4.1 describes the dataset and the preprocessing carried to be out. Section 4.2 shows the results obtained when applying the four clustering validity indices. Finally, Section 4.3 presents the results of the clustering analysis obtained by the k-means.

4.1. Description of the Dataset

As described in the previous section, the first phase is a previous treatment of the raw data. The initial dataset is made up of measurements of electricity consumption with a 15-min frequency taken over six consecutive years. However, these measurements present missing values, which were treated as follows.

Being a time series of 96 elements corresponding to one-day measurements, we find certain empty measurements with zero value. These empty measurements occurred due to point-based errors in the smart meters. In these cases, these zero values precede a very high measurement, well above the average of measurements in that daily interval, corresponding to the accumulation of missing measurements in the previous intervals.

For this reason, these empty values were modified with the mean corresponding to the division of this high value by the number of empty values.

As a result of this cleaning, a RDD variable composed of 8581 rows and 97 columns (the first one with the unique identifier and the remaining ones with electrical consumption measurements) will be analyzed in this work.

The RDD object contains electricity consumption measurements of sensors from the following buildings of Pablo de Olavide University of Sevilla in Spain:

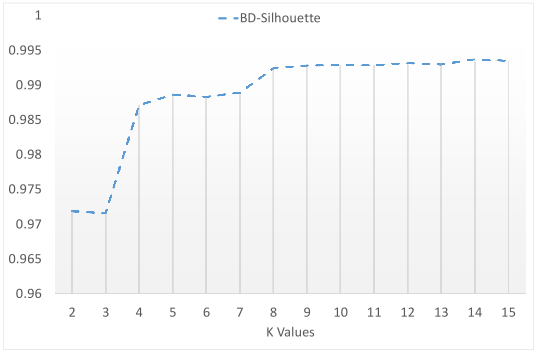
- Building 1—Backup data processing centre (DPC).
- Building 11—Office for professors and classrooms on the ground floor.
- Building 12—Administration services.

- Building 20—Research centre of developmental biology.
- Building 21—Experimental research services.
- Building 42—Old kindergarten (closed since 2010).
- Building 44—Administration services.
- Cafeteria—Cafeteria.

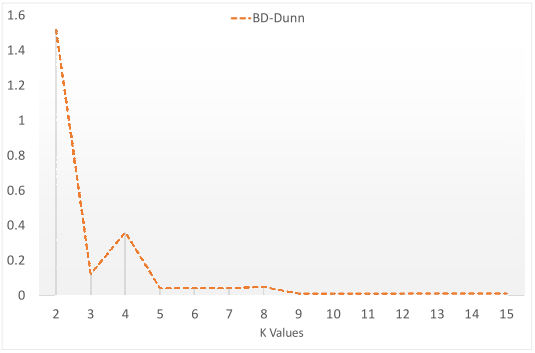
4.2. Cluster Validity Indices Analysis

In this section, the BD-CVIs have been applied to determine the optimal number of clusters to discover useful patterns of electricity consumption in the different buildings of the university.

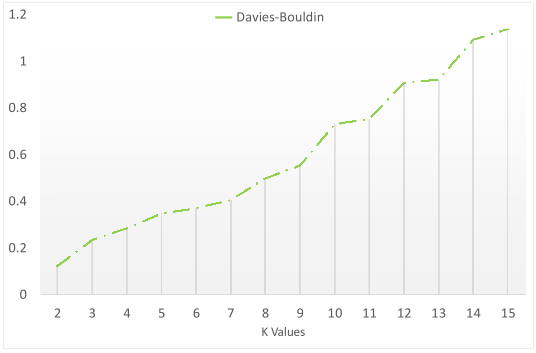
Figure 3 shows the results of the clustering validity indices described in Section 3.2. The results for the BD-Silhouette index are shown in Figure 3a. It can be observed that its curve reaches two local maxima at four and eight. Figure 3b is the BD-Dunn graph and shows local maximum values at four and eight also. The Davies-Bouldin index (Figure 3c) does not show any clear results. However, the curve draws some changes of tendency at 10, 12, and 14, which could be valid results. Figure 3d corresponds to the WSSSE index and draws a stabilization of its values at four and eight. Note that M means millions.



(**a**) BD-Silhouette.



(**b**) BD-Dunn.



(**c**) Davies-Bouldin.

(**d**) WSSSE.

Figure 3. BD-Silhouette, BD-Dunn, Davies-Bouldin, and WSSSE clustering validity indices for *k* values from 2 to 15.

Table 3 shows the results of the BD-CVIs. According to the majority voting method, BD-CVIs suggest that four and eight could be optimal numbers of clusters for the dataset.

Table 3. Majority voting from cluster validation indices (CVIs).

Values	BD-Silhouette	BD-Dunn	Davies-Bouldin	WSSSE
First	**4**	**4**	10	**4**
Second	**8**	**8**	12	**8**
Third	-	-	14	-

4.3. Clustering Results

Clustering results are presented in this Section. As two possible values for the number of clusters have been obtained, this section is divided into two subsections. Sections 4.3.1 and 4.3.2 describe the results when considering four and eight clusters as the optimal number of clusters, respectively.

4.3.1. Analysis of Results: Four Clusters

Table 4 shows the percentage of instances belonging to each cluster. It shows that cluster 1 is the densest, containing 72% of the instances. On the other hand, the consumption centroids for each cluster are displayed in Figure 4. It can be concluded that there are two groups of clusters depending on the consumption level:

- Clusters 2 and 3 with the highest consumptions but with few instances (7% and 4%, respectively).
- Clusters 1 and 4 with the lowest consumptions and the largest percentage of instances (72% and 18%, respectively).

Table 4. Instances along the clusters.

Cluster	Total	Rate
1	6161	72%
2	605	7%
3	311	4%
4	1504	18%

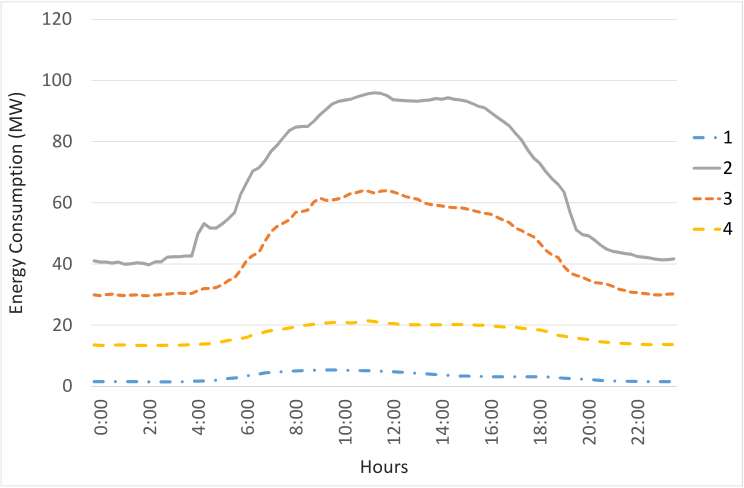
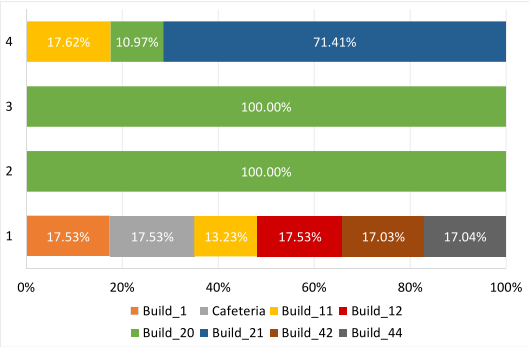


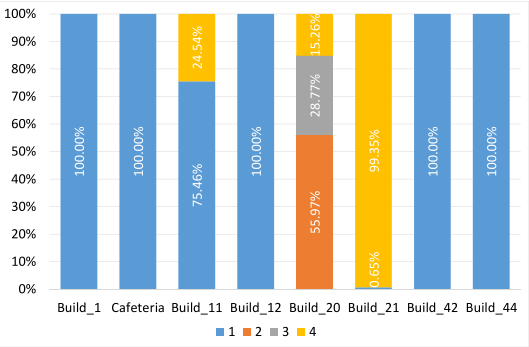
Figure 4. Centroids of the electricity consumption clusters.

Figure 5 shows an analysis of the clusters according to the features buildings, seasons of the year and days of the week. There are two kinds of graphs: Figure 5a,c and e (left side) represent how the clusters are composed of the features, where the bars symbolize the clusters and the colours are the

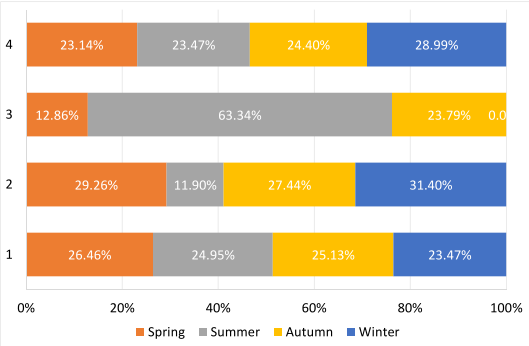
different features. Figure 5b,d and f (right side) represent the presence of the different features in the clusters, where the different features are the columns and the clusters are represented by colours.



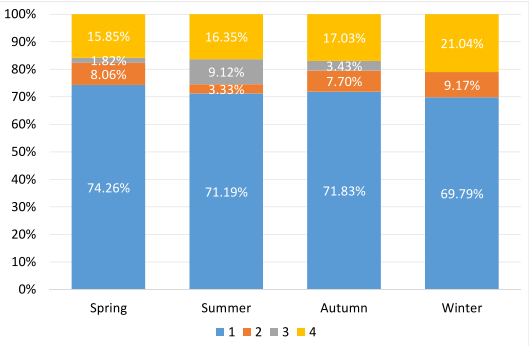
(**a**) Cluster composition according to buildings.



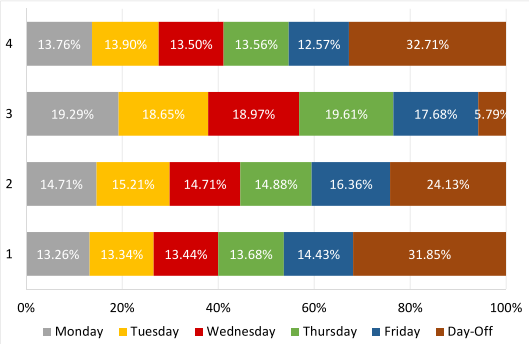
(**b**) Building composition along the clusters.



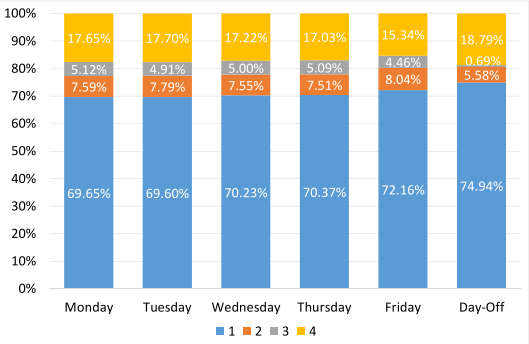
(**c**) Cluster composition according to season of the year.



(**d**) Season of the year composition along the clusters.



(**e**) Cluster composition according to days of the week.



(**f**) Day of the week composition along the clusters.

Figure 5. Cluster analysis depending on buildings, seasons of the year and days of the week.

Figure 5a,b presents the composition of the clusters according to the buildings. Figure 5a shows how the clusters are composed of the different buildings. It can be noticed that clusters 2 and 3 consist of the building 20. Cluster 4 is mainly formed by building 21 -71.41%- and the cluster 1 is equally distributed among all the buildings except buildings 20 and 21. Figure 5b shows the composition of the buildings depending on the clusters. It should be noted that all the buildings, except buildings 20 and 21, have instances in cluster 1, and buildings 1, 12, 42, 44 and cafeteria are just in it.

Figure 5c,d depict a characterization of the clusters according to the feature of the seasons of the year. It is worth noting that cluster 3 is a mainly summer cluster with no instances from winter, and cluster 2 is be the opposite, with few instances of summer and a 31.40% of winter.

Figure 5e,f present the patterns related to the days of the week. It should be highlighted that the percentage of instances is similar during the weekdays. Mainly, the differences exist between the working days and non-working days. Cluster 3 may be considered a working day cluster because the non-working days' instances are just 5.79%. Besides, clusters 1 and 4 have a high rate of instances of non-working days (31.85% and 32.7%, respectively). This fact is consistent with the fact that clusters 1 and 4 were characterized as low-consumption clusters.

Table 5 present the patterns discovered when using four clusters. The characterization of the clusters related to the selected features is summarized as follows:

- Clusters 1 has low consumption and a significant number of instances corresponding to non-working days.
- Cluster 4 has low consumption, and consists of buildings 11 (offices), 20, and 21 (research centres) and instances with a greater presence in non-working days.
- Cluster 2 and 3 have high consumption and both contain building 20, but they are opposites in terms of seasons and days of the week. On the one hand, cluster 2 may be considered a non-summer cluster with a larger number of instances corresponding to non-working days. Although the cluster 2 has a large number of non-working days, the electricity consumption is high because building 20 is dedicated to experimental research. On the other hand, cluster 3 is considered a non-winter cluster, defined by weekdays mainly.

Table 5. Cluster analysis for four clusters.

	Consumption		Buildings			Days	Seasons	
Cluster	**High**	**Low**	**11**	**20**	**21**	**Non-Working Days**	**Non-Summer**	**Non-Winter**
1		✓				✓		
2	✓			✓		✓	✓	
3	✓			✓				✓
4		✓	✓	✓	✓	✓		

4.3.2. Analysis of Results: Eight Clusters

Table 6 shows the number of instances belonging to each cluster after applying k-means with eight clusters. The results show that there are two major clusters, as 39% of the instances belong to clusters 1, 32% to cluster 7 and the rest of the clusters do not reach percentages of 10% each.

Table 6. Instances along the clusters.

Cluster	Total	Rate
1	3333	39%
2	472	6%
3	171	2%
4	274	3%
5	684	8%
6	198	2%
7	2715	32%
8	734	9%

Figures 6 and 7 display the centroids of the clusters representing the average consumptions (in MW), which belong to each cluster within a full day. Figure 6 shows the centroids of all the clusters while Figure 7 shows the centroids with lower consumptions in more detail. Figure 6 reveals that

clusters 2, 3, and 6 have a very high consumption compared to the rest of the clusters. These three clusters have higher consumptions during daylight hours, although the night hours still have a high consumption. Cluster 4 also has a very high consumption, and it remains constant during the day. Clusters 5, 7, and 8 have lower consumptions, which are higher during the daylight hours and much lower during the night. Cluster 1, that contains the largest number of instances, has a consumption close to zero during the entire day.

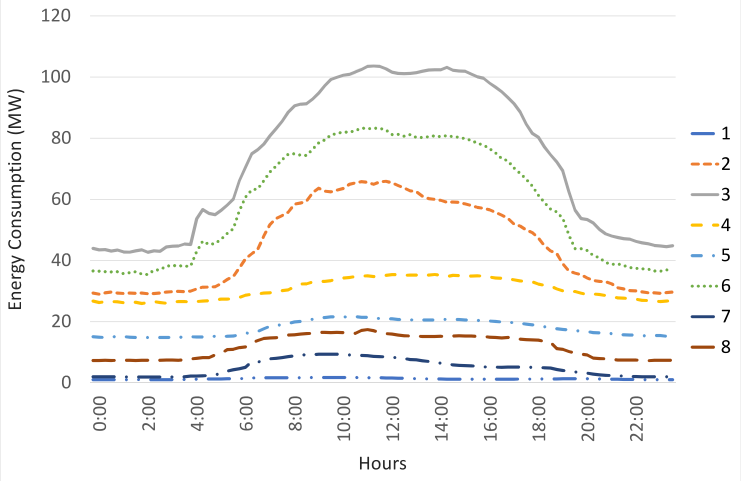


Figure 6. Centroids of the electricity consumption clusters.

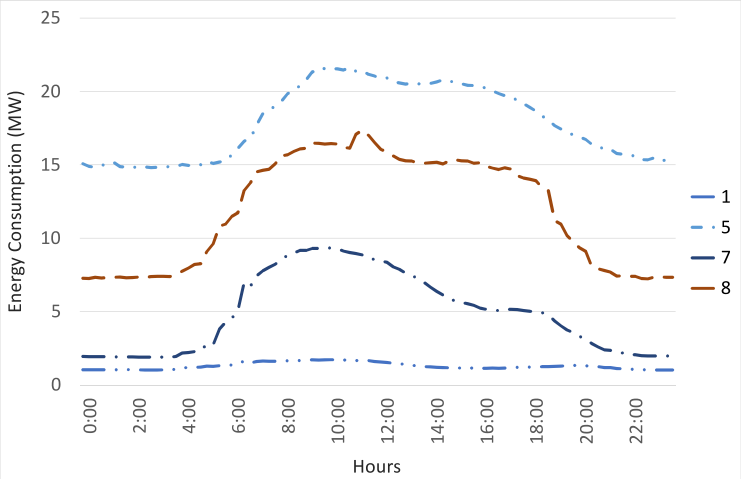
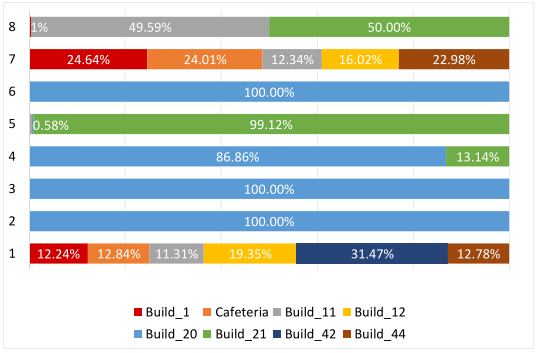


Figure 7. Centroids of the clusters with lower consumptions.

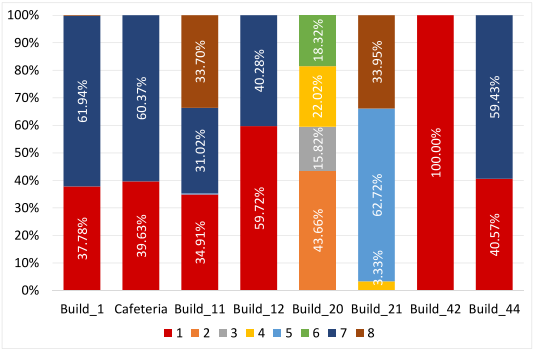
Figure 8 shows an analysis of the clusters obtained when using eight clusters depending on features such as buildings, seasons of the year, and days of the week.

Figure 8a illustrates how the clusters are composed of the buildings in percent. Clusters 2, 3, 4 and 6 are mainly composed of building 20. Besides, cluster 1 is made up of all the buildings except buildings 20 and 21. Cluster 5 consists of the building 21 mainly. The building 21 is also present in cluster 8, that shares half of the instances with building 11. Cluster 7 is formed by instances from all the buildings except buildings 20, 21, and 42. Figure 8b presents the composition of the buildings according to the clusters. It may be highlighted that buildings 1, 11, 12, 44 and the cafeteria belong

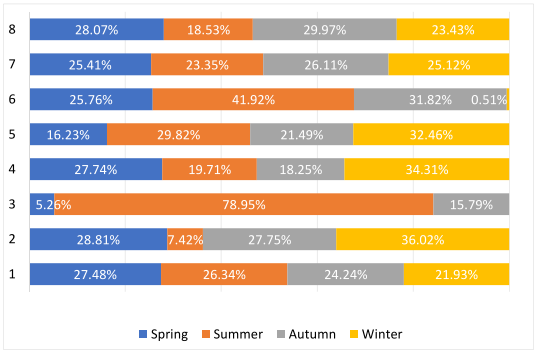
to clusters 1 and 7. Moreover, the buildings 20 and 21 are just the opposite because they belong to different clusters. It is also worth mentioning that the building 42 has all the instances in cluster 1. This is due to building 42 being the old kindergarten closed since 2010, and therefore, this building has no electricity consumption.



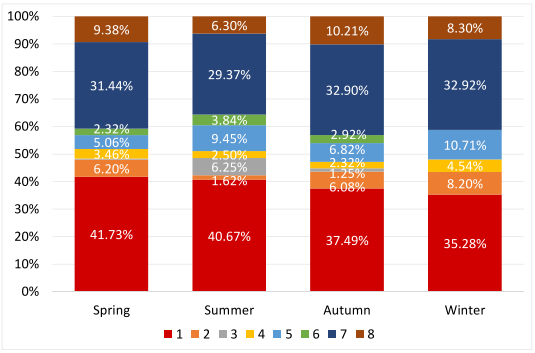
(**a**) Cluster composition according to buildings.



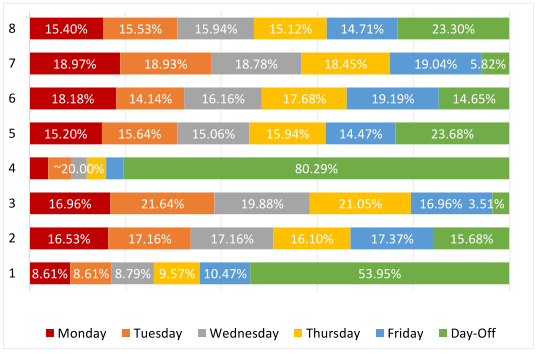
(**b**) Building composition along the clusters.



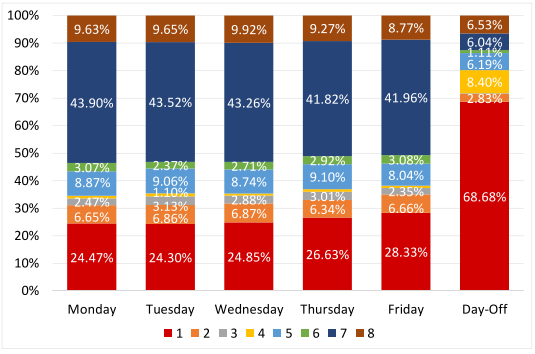
(**c**) Cluster composition according to seasons of the year.



(**d**) Season of the year composition along the clusters.



(**e**) Cluster composition according to days of the week.



(**f**) Day of the week composition along the clusters.

Figure 8. Cluster analysis depending on buildings, seasons of the year, and days of the week.

Figure 8c presents how the clusters are composed of seasons of the year. It can be appreciated that clusters generally have instances equally distributed over the seasons with some exceptions. For instance, cluster 2 has instances during all the seasons but summer, and the opposite situation is found in cluster 3, which has more instances corresponding to summer days. Furthermore, clusters 2, 4, and 5 have a percentage of instances slightly higher in winter: 36.02%, 34.31%, and 32.46%, respectively.

It worth mentioning that cluster 6 is composed of non-winter instances as just a 0.51% of instances correspond to winter and 41.92% to summer.

Figure 8e shows the distribution of the days of the week depending on the clusters. It is worth noting that clusters 1 and 4 are mainly composed of non-working days, just the opposite to clusters 3 and 7 that only have 3.51% and 5.82% of day-off instances, respectively. Figure 8f presents the percentage of instances of each cluster composing each type of day. It can be emphasized that days off are mainly composed of cluster 1 instances, while the rest of days are mostly formed by cluster 7.

Table 7 provides a characterization of the clusters obtained when using 8 clusters by means of the features analysed A summary is described below:

- Cluster 1 contains the instances with the lowest consumption and that are constant throughout the day. It is composed of all the buildings except buildings 20 and 21 (research centres). The instances are mostly non-working days and they are distributed uniformly over all seasons of the year.
- Clusters 2, 3, 4, and 6 are composed of building 20. These clusters contain the highest consumption during daylight hours. Clusters 2 and 6 include instances from all the days of the week, while clusters 3 and 4 just have instances from working days and non-working days, respectively. Most of the instances of the cluster 2 are non-summer instances, and cluster 3 is just the opposite because it includes summer instances mainly.
- Cluster 5 is composed of building 21. It is characterized by a low consumption which is higher during daylight hours. In addition, it contains instances of all the days of the week but slightly more for non-working days.
- Cluster 7 consists of all the buildings, except 20, 21, and 42. It represents a low consumption higher during daylight hours and working days.
- Cluster 8 is formed by the buildings 11 (offices) and 21. It represents low consumption but higher during daylight hours and non-working days.

Table 7. Cluster analysis for eight clusters.

	Consumption			**Days**		**Seasons**			**Buildings**		
Cluster	**High**	**Low**	**Diurnal**	**Working Days**	**Non-Working Days**	**Non-Summer**	Summer	**Non-Winter**	**11**	**20**	**21**
1		✓			✓				✓		
2	✓		✓			✓				✓	
3	✓		✓	✓			✓			✓	
4	✓				✓					✓	
5		✓	✓								✓
6	✓		✓					✓		✓	
7		✓	✓	✓					✓		
8		✓	✓		✓				✓		✓

5. Execution Times

This section provides the computing times using different synthetic big data to evaluate the scalability of the proposed methodology. To this end, the set of the electricity consumptions from the eight buildings located on the university campus has been exponentially increased, with the aim of simulating a neighbourhood, a town, a city or a metropolis.

Let us remember that the data comes from smart meters every 15 min for six years for eight buildings. Mathematical operations defined in set theory such as union, distinct, or join, which are supported by Spark technology, were used in order to transform datasets into exponentially bigger ones. In particular, the input datasets were generated by means of union operations from the original ones.

Table 8 shows computing times obtained by the proposed methodology using synthetic datasets for two different hardware configurations. Each row describes information about each of the generated datasets such as the number of buildings, total number of instances, size of the file and runtimes measured in hours. $Time_1$ shows runtimes using a five-node cluster with 8 CPUs and 15 GB RAM,

and the $Time_2$ column is the second hardware scenario where there is a five-node cluster composed of 16 CPUs and 30 GB RAM.

Table 8. Computing times (in hours) using synthetic big data for two different hardware configurations.

Buildings	Instances	File Size	$Time_1$	$Time_2$
16	17,162	10.3 MB	0.0015	0.0015
32	34,324	20.5 MB	0.0015	0.0016
64	68,648	41.2 MB	0.0015	0.0014
128	137,296	82.4 MB	0.0014	0.0015
256	274,592	190.1 MB	0.0018	0.0017
512	549,184	380.9 MB	0.0021	0.0015
1024	1,098,368	744.1 MB	0.0023	0.0022
2048	2,196,736	1.45 GB	0.0037	0.0020
4096	4,393,472	2.91 GB	0.0067	0.0023
8192	8,786,944	5.81 GB	0.0094	0.0054
16,384	17,573,888	11.63 GB	0.0156	0.0091
32,768	35,147,776	23.26 GB	0.7078	0.0162
65,536	70,295,552	46.52 GB	3.8555	0.0995
131,072	140,591,104	93.03 GB	5.2325	1.1985

On the one hand, in the first hardware scenario, execution times are negligible up until the dataset of 11.63 GB composed of 16,384 buildings. The largest dataset with 131,072 buildings, big data that could represent a big metropolis, had a time of 5.2325 h. On the other hand, the second hardware configuration keeps slight runtimes, at 0.0995 h, up until the dataset with 65,536 buildings, with 1.1985 h for the largest dataset. It should be highlighted that the first configuration obtained reasonable times, but using a more powerful hardware configuration, times are reduced considerably. For the largest dataset, execution time has been reduced up to 5 times and about 40 times for the dataset with 65,536 buildings.

Computational times of the different processes in the methodology are proportional in the two hardware configurations. Taking into account all the phases of the methodology, obtaining the optimum number of clusters is the process that takes the longest, occupying 72% of the total time. This is because it is an iterative process in which k-means is launched along the indices n times, where n is the maximum number of clusters we could assume. Within this process, k-means takes 85%, and the rest of the time is used to calculate the values of the clustering validity indices. The next phases that take longer are clustering and preprocessing analysis, lasting 13% and 11%, respectively. Finally, the calculation of the k-means takes the shortest time, since it simply launches the algorithm with already preprocessed data and an optimal number of clusters.

Figure 9 graphically shows runtimes in hours when increasing the number of buildings for the two different hardware configurations. As it can be noticed, both runtimes are similar using datasets with less than 20,000 buildings, but the difference between both configurations is quite remarkable for 60,000 buildings. In particular, $Time_1$ was 40 times larger than $Time_2$. As it can be seen, results show that techniques described in this paper can be applied to optimize the electricity consumption of a smart city within a reasonable time.

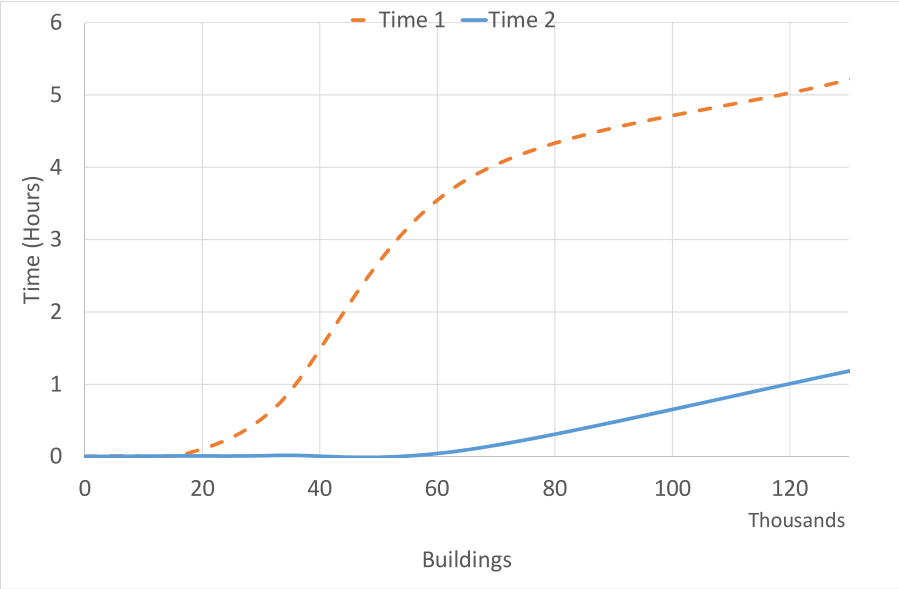


Figure 9. A runtime comparison between the two different hardware configurations.

6. Conclusions

A detailed understanding of energy consumption patterns of buildings is essential for smart cities. On the one hand, electricity companies can improve the pricing policies and offer customized packages for certain types of communities. On the other hand, public administrations can optimize resources by contracting certain hourly rates of discrimination offered by the main electricity companies in the countries. A joint application of the methodology proposed in this work by energy companies and public administrations can bring benefits to the community as a whole since energy saving is essential to reduce the impact on climate change and promote sustainable development. In this context, we propose a work methodology to detect patterns from big data time series, as this type of data is generated by modern smart cities through the increasingly common smart meters.

In this paper, a model based on the k-means algorithm was designed for this purpose using the distributed computing advantages of Apache Spark. Firstly, a study of four CVIs optimized for parallelization—the DB-Dunn, DB-Silhouette, Davies-Bouldin and WSSSE indices—was carried out. From these indices, a majority voting strategy was applied in order to choose the optimal number of clusters. This study returned two possible values for the number of clusters and an in-depth analysis of the patterns for both cases was performed.

Next, patterns were characterized according to the building, type of consumption (high, low, daytime or constant), the season of the year, and day of the week (including days off). A valuable interpretation of the patterns obtained has been provided. Namely, the consumption behaviour of buildings depends mainly on their characteristics (administration buildings, research centres, classrooms or leisure facilities) and the hours during the day which they are used. In addition, it has been shown that there is a strong relationship between temperature and consumption, and a high impact of holiday periods in the academic calendar.

Finally, several synthetic datasets were generated from the original dataset. These datasets were used to measure computing times required to discover patterns using the proposed methodology. Results showed a linear relationship between runtimes and size of datasets. In fact, the execution time for the largest dataset considering big data is less than 4 h. Thus, in the hypothetical case of obtaining a dataset with six-year measurements for 65,536 buildings, the runtime is computationally suitable.

Future work will focus on two aspects: firstly, to discover consumption patterns in big data using other additional variables (such as price or type of consumer), and secondly, the prediction of electricity consumption from big data using distributed technology such as Apache Spark. The complete methodology proposed in this paper allows us to lay the foundations for the use of different prediction algorithms, once the original data set has been clustered. In this sense, algorithms in distributed technology are being developed to obtain predictions with high accuracy.

These two approaches will support the economic and political decision making of different public administrations, as well as the personalization of products by private organizations (energy companies, for example), increasingly involved in tracking their resources to obtain valuable information in the context of smart cities.

Acknowledgments: This work has been supported by the Spanish Ministry of Economy and Competitiveness and Junta de Andalucía under projects TIN2014-55894-C2-R, TIN2017-88209-C2-R and P12-TIC-1728. J.M.L.-R. holds a FPI scholarship from the Spanish Ministry of Economy and Competitiveness.

Author Contributions: R.P.-C. and J.M.L.-R. implemented the methodology and drafted the manuscript; J.C.R. conceived and designed the experiments; A.T. and F.M.-Á. participated in the elaboration of the manuscript; all authors read, edited and approved the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MLlib	Machine Learning Library
CVI	Cluster Validity Index
BD-CVI	Big Data Cluster Validity Index
RDD	Resilient Distributed Dataset
DPC	Data Processing Centre
AWS	Amazon Web Services

References

1. Nuaimi, E.A.; Neyadi, H.A.; Mohamed, N.; Al-Jaroodi, J. Applications of big data to smart cities. *J. Internet Ser. Appl.* **2015**, *6*, 1–15.
2. Gungor, V.C.; Sahin, D.; Kocak, T.; Ergut, S.; Buccella, C.; Cecati, C.; Hancke, G.P. Smart Grid Technologies: Communication Technologies and Standards. *IEEE Trans. Ind. Inf.* **2011**, *7*, 529–539.
3. Fernández, A.; del Río, S.; López, V.; Bawakid, A.; del Jesús, M.J.; Benítez, J.M.; Herrera, F. Big Data with Cloud Computing: An insight on the computing environment, MapReduce, and programming frameworks. *Wiley Interdiscip. Rew. Data Min. Knowl. Discov.* **2014**, *4*, 380–409.
4. Orgaz, G.B.; Jung, J.J.; Camacho, D. Social big data: Recent achievements and new challenges. *Inf. Fusion* **2016**, *28*, 45–59.
5. Dean, J.; Ghemawat, S. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM* **2008**, *51*, 107–113.
6. Zaharia, M.; Chowdhury, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Spark: Cluster Computing with Working Sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*; HotCloud'10; USENIX Association: Berkeley, CA, USA, 2010; p. 10.
7. Zaharia, M.; Chowdhury, M.; Das, T.; Dave, A.; Ma, J.; McCauley, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-memory Cluster Computing. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*; NSDI'12; USENIX Association: Berkeley, CA, USA, 2012; p. 2.
8. Meng, X.; Bradley, J.; Yavuz, B.; Sparks, E.; Venkataraman, S.; Liu, D.; Freeman, J.; Tsai, D.; Amde, M.; Owen, S.; et al. MLlib: Machine Learning in Apache Spark. *J. Mach. Learn. Res.* **2016**, *17*, 1–7.
9. Bahmani, B.; Moseley, B.; Vattani, A.; Kumar, R.; Vassilvitskii, S. Scalable K-means++. *Proc. VLDB Endow.* **2012**, *5*, 622–633.

10. Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; PéRez, J.M.; Perona, I.N. An Extensive Comparative Study of Cluster Validity Indices. *Pattern Recogn.* **2013**, *46*, 243–256.
11. Luna-Romera, J.M.; García-Gutiérrez, J.; Martínez-Ballesteros, M.; Santos, J.C.R. An approach to validity indices for clustering techniques in Big Data. *Prog. Artif. Intell.* **2017**, *7*, 1–14.
12. Martínez-Álvarez, F.; Troncoso, A.; Riquelme, J.C.; Ruiz, J.S.A. Energy Time Series Forecasting Based on Pattern Sequence Similarity. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 1230–1243.
13. Tuballa, M.L.; Abundo, M.L. A review of the development of Smart Grid technologies. *Renew. Sustain. Energy Rev.* **2016**, *59*, 710–725.
14. Calvillo, C.; Sánchez-Miralles, A.; Villar, J. Energy management and planning in smart cities. *Renew. Sustain. Energy Rev.* **2016**, *55*, 273–287.
15. Sun, Y.; Song, H.; Jara, A.J.; Bie, R. Internet of Things and Big Data Analytics for Smart and Connected Communities. *IEEE Access* **2016**, *4*, 766–773.
16. Xu, J.; Zhang, R. CoMP Meets Smart Grid: A New Communication and Energy Cooperation Paradigm. *IEEE Trans. Vehicular Technol.* **2015**, *64*, 2476–2488.
17. Wijk, J.J.V.; Selow, E.R.V. Cluster and calendar based visualization of time series data. In Proceedings of the IEEE Symposium on Information Visualization, San Francisco, CA, USA, 24–29 October 1999; pp. 4–9.
18. Martínez-Álvarez, F.; Troncoso, A.; Riquelme, J.C.; Riquelme, J.M. Partitioning-Clustering Techniques Applied to the Electricity Price Time Series. In Proceedings of the Intelligent Data Engineering and Automated Learning, Birmingham, UK, 16–19 December 2007; pp. 990–999.
19. Martínez-Álvarez, F.; Troncoso, A.; Riquelme, J.C.; Riquelme, J.M. Discovering patterns in electricity price using clustering techniques. In Proceedings of the International Conference on Renewable Energy and Power Quality, Sevilla, Spain, 28–30 march 2007; pp. 245–252.
20. Keyno, H.S.; Ghaderi, F.; Azade, A.; Razmi, J. Forecasting electricity consumption by clustering data in order to decline the periodic variable's affects and simplification the pattern. *Energy Convers. Manag.* **2009**, *50*, 829–836.
21. Hernández, L.; Baladrón, C.; Aguiar, J.M.; Carro, B.; Sánchez-Esguevillas, A. Classification and Clustering of Electricity Demand Patterns in Industrial Parks. *Energies* **2012**, *5*, 5215–5228.
22. Fahad, A.; Alshatri, N.; Tari, Z.; Alamri, A.; Khalil, I.; Zomaya, A.Y.; Foufou, S.; Bouras, A. A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Trans. Emerg. Top. Comput.* **2014**, *2*, 267–279.
23. Ding, R.; Wang, Q.; Wang, Q.; Dang, Y.; Fu, Q.; Zhang, H.; Zhang, D.; Ding, J. YADING: Fast Clustering of Large-Scale Time Series Data. *Proc. Very Large Data Bases* **2015**, *8*, 473–484.
24. Rakthanmanon, T.; Campana, B.; Mueen, A.; Batista, G.; Westover, B.; Zhu, Q.; Zakaria, J.; Keogh, E. Addressing Big Data Time Series: Mining Trillions of Time Series Subsequences Under Dynamic Time Warping. *ACM Trans. Knowl. Discov. Data* **2013**, *7*, 1–31.
25. Zhao, W.; Ma, H.; He, Q. Parallel K-Means Clustering Based on MapReduce. In *Cloud Computing*; Jaatun, M.G., Zhao, G., Rong, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 674–679.
26. Capó, M.; Pérez, A.; Lozano, J.A. An Efficient Approximation to the K-means Clustering for Massive Data. *Know.-Based Syst.* **2017**, *117*, 56–69.
27. Melzi, F.N.; Same, A.; Zayani, M.H.; Oukhellou, L. A Dedicated Mixture Model for Clustering Smart Meter Data: Identification and Analysis of Electricity Consumption Behaviors. *Energies* **2017**, *10*, 1–21.
28. Deb, C.; Zhang, F.; Yang, J.; Lee, S.E.; Shah, K.W. A review on time series forecasting techniques for building energy consumption. *Renew. Sustain. Energy Rev.* **2017**, *74*, 902–924.
29. Singh, S.; Yassine, A. Big Data Mining of Energy Time Series for Behavioral Analytics and Energy Consumption Forecasting. *Energies* **2018**, *11*, 452.
30. Li, C.; Ding, Z.; Zhao, D.; Yi, J.; Zhang, G. Building Energy Consumption Prediction: An Extreme Deep Learning Approach. *Energies* **2017**, *10*, 1525.
31. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *1*, 224–227.
32. Spark, A. Clustering—RDD-Based API—Spark 2.2.0 Documentation. 2017. Available online: https://spark.apache.org/docs/2.2.0/mllib-clustering.html#k-means (accessed on 20 December 2017).

33. Ketchen, D.J.; Shook, C.L. The Application Of Cluster Analysis In Strategic Management Research: An Analysis And Critique. *Strateg. Manag. J.* **1996**, *17*, 441–458.
34. Koprinska, I.; Rana, M.; Troncoso, A.; Martínez-Álvarez, F. Combining pattern sequence similarity with neural networks for forecasting electricity demand time series. In Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 4–9 August 2013; pp. 1–8.

Capítulo 7

Analysis of the evolution of the Spanish labour market through unsupervised learning

Resumen

En este artículo se analiza el mercado laboral español aplicando técnicas de aprendizaje no supervisado con tecnología de Big Data. El objetivo de este análisis es el de descubrir como se organiza el mercado laboral teniendo en cuenta los diferentes tipos de trabajadores y trabajos que existen en el territorio nacional. El análisis se realiza a dos periodos laborales con mucha repercusión económica: 2011-2013, años de plena crisis económica; y 2014-2016 que podría ser considerado un periodo de recuperación económica. Los datos usados provienen del Ministerio de Trabajo, Migraciones y Seguridad Social, y corresponden a 1.9 y 2.3 millones de contrataciones respectivamente. Para llevar a cabo el análisis se han usado dos algoritmos de clustering diferentes, k-means y average linkage, y dos tecnologías software diferentes, Stata y Apache Spark. La investigación de este estudio presenta los efectos de la crisis económica en el mercado laboral español. Los resultados indican que ha habido transformaciones en el mercado, lo cual ha repercutido en la fisionomía de algunos nichos de trabajo, sin embargo, se presentan diferentes *clusters* de trabajos que han perdurado en el tiempo a pesar de la crisis. Además, el artículo hace una comparativa de los resultados ofrecidos por k-means y average linkage, y muestran grandes similitudes entre ellos, con un ratio entre el 66 % y el 98 % que varía en función del número de *clusters* que tengamos en cuenta. Estos resultados pueden llegar a apoyar las decisiones económicas y políticas de las diferentes administraciones públicas, así como mejorar las futuras políticas de empleo.

- Estado: 2ª ronda de revisión en IEEE Access (IEEE)

- Índice de Impacto (JCR 2018): 4.098
- Área de Conocimiento:
  - Engineering, electrical & electronics. Ranking 52/265 - Q1
  - Telecommunications. Ranking 19/88 - Q1
  - Computer Science, Information Systems. Ranking 24/155 - Q1

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.DOI

Analysis of the evolution of the Spanish labour market through unsupervised learning

J. M. LUNA-ROMERA[1*], F. NÚÑEZ-HERNÁNDEZ[2*], M. MARTÍNEZ-BALLESTEROS[1], J. C.RIQUELME[1], C. USABIAGA IBÁÑEZ[3]

[1]Division of Computer Science, University of Seville, Seville, Spain ({jmluna,jfabregas,riquelme}@us.es).
[2]Department of Industrial Organisation, University of Seville, Seville, Spain (fnunez@us.es).
[3]Department of Economics, Pablo de Olavide University, Seville, Spain (cusaiba@upo.es).
[*]These authors contributed equally.

This work has been supported by the Spanish Ministry of Economy and Competitiveness under project TIN2014-55894-C2-R. J.M. Luna-Romera holds a FPI scholarship from the Spanish Ministry of Economy and Competitiveness. Fernando Núñez-Hernández and Carlos Usabiaga acknowledge the funding from the Spanish Ministry of Economics, Industry and Competitiveness (ECO2017-86780-R Project) and the Andalusian Government (SEJ-513 PAIDI Research Group).

ABSTRACT Unemployment in Spain is one of the biggest concerns of its inhabitants. Its unemployment rate is the second highest in the European Union, and in the second quarter of 2018 there is a 15.2% unemployment rate, some 3.4 million unemployed. Construction is one of the activity sectors that have suffered the most from the economic crisis. In addition, the economic crisis affected in different ways to the labour market in terms of occupation level or location. The aim of this paper is to discover how the labour market is organised taking into account the jobs that workers get during two periods: 2011-2013, which corresponds to the economic crisis period, and 2014-2016, which was a period of economic recovery. The data used are official records of the Spanish administration corresponding to 1.9 and 2.4 million job placements, respectively. The labour market was analysed by applying unsupervised machine learning techniques to obtain a clear and structured information on the employment generation process and the underlying labour mobility. We have applied two clustering methods with two different technologies, and the results indicate that there were some movements in the Spanish labour market which have changed the physiognomy of some of the jobs. The analysis reveals the changes in the labour market: the crisis forces greater geographical mobility and favours the subsequent emergence of new job sources. Nevertheless, there still exist some clusters that remain stable despite the crisis. We may conclude that we have achieved a characterisation of some important groups of workers in Spain. The methodology used, being supported by Big Data techniques, would serve to analyse any alternative job market.

INDEX TERMS Labour market, Cluster analysis, Labour mobility, Big data.

I. INTRODUCTION

The unemployment rate in Spain is the second highest (15%) among the countries of the European Union after Greece (19%). In recent years the unemployment rate has doubled the European Union average, and the rates are even worse if we focus on youth unemployment, which in 2014 reached 57.9% [1]. Currently, the unemployment rate has a tendency to decline, but it is the cause that most worries to the Spanish, followed by corruption and economic problems [2].

Over recent decades, the Spanish economy has been rooted on a traditional production model based on sectors such as construction and tourism, which, at the end of the last boom period, accounted for more than a quarter of the national production. In 2008, just at the beginning of the recent economic crisis, the construction sector was around 15% of Spanish GDP -and in addition we would have to take into account the important linked activities-, and tourism was around 11% (in general, the Spanish economy is characterised by an important tertiary bias). Thereby, the collapse of Spain's construction sector -jointly with several related activities- after the bursting of the real estate-financial bubble has beaten records in increasing unemployment at a speed never seen

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

before -the unemployment rate rose from less than 10% to more than 25% in just a few years. Several million jobs were destroyed during the recent economic crisis, going from 20.6 million employed at the first quarter of 2008 to 16.9 million at the first quarter of 2014 -around 1.7 million lost jobs were in the construction sector. In those years the long-term unemployment problem also regained strength. Fortunately, the labour market figures have moderately improved in the most recent years, although with problems in the quality of the jobs generated.

In the labour market, workers looking for jobs and vacant jobs offered by firms are heterogeneous in many aspects: skills, geographical location, gender, age, payment, etc. These heterogeneities lead to the concept of mismatch: "Mismatch is an empirical concept that measures the degree of heterogeneity in the labour market across a number of dimensions, usually restricted to skills, industrial sector, and location" [3, p. 399].

In this paper, employment data in Spain is processed in order to characterise and to identify groups at the Spanish labour market in order to analyse the evolution that has occurred during and after the crisis. For that reason, we have applied unsupervised machine learning techniques which allow us to discover knowledge from data with just its intrinsic information. In this context, there exists the clustering analysis, that is defined in [4] as the process of partitioning a set of data objects into subsets, where each subset is a cluster; objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. Thereby, clustering is useful in that it can lead to the discovery of previously unknown groups within the data. A clustering analysis is proposed in this study in order to account for the role of heterogeneities in the matching process of the Spanish labour market.

Specifically, we have applied two different clustering algorithms: firstly, partitional and well-known k-means algorithm [5]; secondly, hierarchical clustering with average linkage. One of the main difficulties of cluster analysis is finding the optimal number of clusters. In the k-means algorithm is a prerequisite while in the hierarchical proposal can be decided a posteriori. In this work we have applied both internal and external validation indices to decide the number of clusters in the k-means algorithm, while with average linkage we have opted for a choice in two stages: first we have chosen k based on an internal index and a subsequent refinement based on minimising k with maximum representativeness.

Both clustering techniques have been applied to data from the Spanish labour market in two different economic periods: 2011-2013, which corresponds to an economic crisis period in Spain, and 2014-2016, which has been a period of economic recovery.

The application of both clustering methods to those different economic periods gives four results that are analysed and compared among them. The objective is discuss the evolution of the Spanish labour market over these years of significant economic changes. The main contribution of this article from the labour perspective is to apply unsupervised machine learning techniques to obtain a clearer and more structured information on the employment generation process and the underlying labour mobility. This information tool, based on the recent labour matching flow, should allow the authorities to orientate, geographically and occupationally, the worker's search.

The methodology applied in this work is based on Big Data implementations that would allow the analysis performed to be extended to any volume of data regardless of the length of time period analysed or the size of the labour market of a country or international organisations.

The rest of the paper is organised as follows: Section II presents the related works from the literature. Section III establishes the applied methodology including the complete process that is carried out. Section IV details the results, including those that are accomplished by k-means and by average linkage. Finally, Section V summarises the conclusions of our study.

II. RELATED WORK

Data mining is one of the most successful fields of statistics and computer science that uses machine learning, artificial intelligence, statistics and database systems to analyse information in order to discover implicit, new, and potentially useful knowledge from data. Machine learning is the area of artificial intelligence that aims at developing systems that learn automatically and relies on finding patterns and relationships within the data, known as training data, to create models, that is, abstract representations of reality [6]. The training data is composed by a set of examples and each example is characterised by a set of features.

Machine learning tasks are mainly classified into supervised and unsupervised learning. In supervised learning, a mathematical model is created from a set of data that contains the input values and needs a ground truth or prior knowledge of what the output values should be. The most common types of supervised learning are classification (limited set of values for the outputs) and regression (continuous outputs) algorithms. On the contrary, the data only contains input values but does not require labelled output values in unsupervised learning. This kind of algorithms aims to infer the underlying structure or distribution in the data. They can identify patterns or relationships between examples or between features depending on whether they are clustering or association rules algorithms, respectively.

Clustering is one of the most used unsupervised machine learning techniques. Clustering groups the data in clusters so that those data that belong to the same cluster share similar features or attributes, and that data is dissimilar to those in other clusters. The similarity of the data is normally given by how close they are in space, taking into account a distance function [4].

There are many clustering methods in the literature, and there are some works that classify them by some criteria [4, 7, 8]. In this paper we are going to focus on partitional and hierarchical clustering methods. The basic idea of clustering

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

based on partitioning is to divide the data into k groups such that the elements which belong to the same group are more similar than the elements from different groups as can be observed in Figure 1. Many partitioning methods form clusters based on distances, so that k clusters are initially assigned, and the object clusters are iteratively changed until a solution where each object is in its nearest cluster is found [9], such as the well-known k-means algorithm [5].

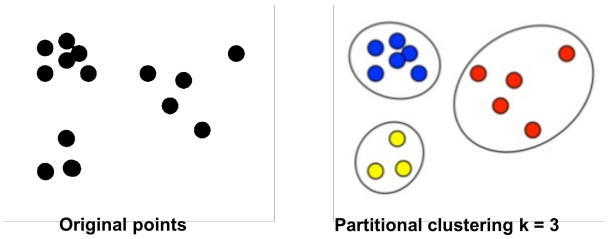


FIGURE 1: Example of clustering based on partitional methods.

Hierarchical methods create a hierarchical decomposition of the given set of data objects. They can be considered as agglomerative or bottom-up clustering methods if the hierarchy is built assigning each object to its own cluster and then, the most similar clusters are iteratively joined until only a single cluster is left. On the contrary, they can be denoted as divisive or top-down clustering methods if the clusters are created in reverse manner. Thus, all objects are assigned to a single cluster which is recursively split until there is one cluster for each object [10, 11]. The average linkage hierarchical clustering is one of the commonly used hierarchical algorithms where the distance between two clusters is determined by the average distance between each point in one cluster to every point in the other cluster [11].

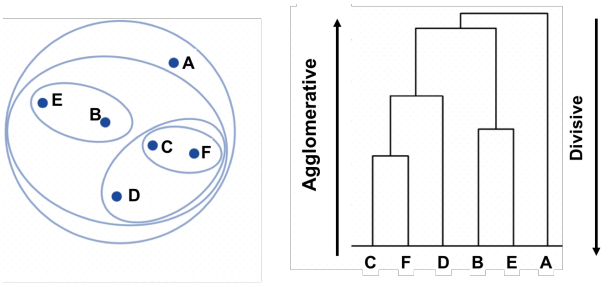


FIGURE 2: Example of clustering based on hierarchical methods.

These methods require a number of clusters into which the data is going to be partitioned. The main problem is that the optimal number of clusters is not known until the clustering is done. This task has been handled in the literature in diverse works [12, 13] establishing the named clustering validity indices (CVI), which are metrics that measure the quality of the clustering. There exists a taxonomy in the literature that distinguishes between two kinds of CVI: internal indices, which measure the quality of the clustering results according to the distance between the clusters, and the compactness of the objects that belong to the same cluster; and external indices, which measure the quality of the clustering solution through an external indicator of the object distinguishes such as the class.

This paper applies clustering methods to the MCVL information on registered job matches in the Spanish labour market. The nature of our data, with information about jobs and workers having productive matches, links up our work directly with the theoretical concept of the aggregate matching function. This function represents the labour matching process without the need to make explicit the heterogeneities and labour frictions. Instead of representing them specifically according to their origin and their type, heterogeneities and labour frictions are implicitly introduced into an aggregate function that relates the flow of job placements in each period with the levels and inflows of vacancies and job seekers (mainly unemployed seekers). There is an extensive literature (theoretical and empirical) on job search and labour matching processes and, in particular, on the aggregate matching function [3, 14, 15, 16, 17]. It is important to note that the matching function assumes that workers and jobs are heterogeneous but omits to make those heterogeneities explicit. Without heterogeneities (zero mismatch), the matching function would not exist and jobs and workers would match instantaneously [3, 18, 19, 20].

Considerable work has been carried out in the literature to open the 'black box' of the matching process and to make explicit the heterogeneities hidden in the matching function. Island models can be found in [21, 22]; urn-ball models in [3]; the taxicab model in [23];queuing models in [24, 25]; stock-flow models in [26, 27]; and mismatch models in [20]. As a rule, in all these models, workers and jobs are divided into parts (local labour markets, locations, islands, queues, worker-job pairs acceptable or unacceptable to match productively, stock (old)-flow (new) workers and jobs), which are then treated as if each part were homogeneous. Therefore, it is assumed that the heterogeneities of workers and jobs are the reason that the labour market is segmented. Features such as skills, location, age, sex, etc., make certain jobs only suitable for certain workers -there exists evidence of labour market segmentation in the Spanish economy, based primarily on skills and location [28, 29].

The existence of homogeneous groups of workers (and jobs) in a segmented market gives validity to the use of clustering techniques to analyse the matching process in the labour market. Since highly detailed division of the MCVL data in workers or job categories results in a very large number of units, which may be difficult to understand and analyse, we use a clustering methodology, based on a similarity measure, to obtain larger homogeneous groups (clusters) and a better overview of the structure of the labour market [30, 31] which is compatible with the existing theories on labour matching. Cluster analysis enables, as far as possible, subjective or 'a priori' similarity criteria to be avoided -

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

grouping provinces in greater administrative regions, for instance-. Instead, we look for a similarity criterion that is consistent with the search and matching theories applied to labour economics. In this sense, we consider that worker (job) categories are more similar the more they resemble in the way they match with job (worker) categories; as we shall see, the Manhattan distance is compatible with this idea of similarity in the matching process. It should be highlighted that we have used Manhattan distance based on the works from [32, 33], which used a variant of Manhattan whose values are in the interval $[0, 1]$. Manhattan distance between two worker categories W_i and W_j is defined as:

$$d(W_i, W_j) = \sum_{z=1}^{n} |W_{i_z} - W_{j_z}| \tag{1}$$

where n is the total of job categories.

Our study follows the research line of [32, 33] consisting in applying cluster analysis to labour matching data. Other studies have introduced matching data in the analysis of labour clusters. For example, [34] analyses the labour mobility between clusters in Stockholm taking as reference the information and communications technology (ICT) cluster. For these authors, a labour cluster is not simply a large number of firms that belong to the same industrial sector, but a set of complementary and interlinked firms and institutions that have developed a shared consciousness and identity as an industrial cluster and system. In [35] a computer programme that identifies localised mobility clusters in Sweden is developed, the clusters are based on the flows of job movers between workplaces. According to these authors, traditional pecuniary externalities have to be combined with technological and knowledge externalities, coming from the exchange of labour between firms, in order to implement a complete cluster analysis. In this line, the study from [36] used a large Portuguese employer-employee panel-data set to study Marshall's hypothesis that industrial agglomeration improves the quality of firm-worker matching. For these authors, the formation of industrial clusters produces external scale economies, since it increases three intangibles: the potential for more extensive interaction between suppliers and buyers, the firms' ability to capture industry-specific knowledge spillovers resulting from the close proximity of similar firms, and the number of available labour skills and the quality of firm-worker matching. Other articles have analysed labour clusters but without using matching data. For instance, in [37] is studied, for the UK, whether or not different empirical techniques produce identical or similar results in classifying labour markets into homogeneous entities; obtaining some evidence of segmentation in the labour market. The study in [38] worked with a micro-database on workers, for the region of Aragón in Spain, which provides information, among other variables, about where the worker lives and where the worker works. The objective of these authors is to identify local labour markets (clusters) in which a large proportion of the workers both live and work. Meanwhile, following a macroeconomic approach, these works [39, 40, 41] apply cluster analysis to the Spanish, the European and the German labour market respectively, all of them from a regional perspective. The first authors show that high and low unemployment Spanish regions have similar responses to regional employment shocks in the short-run, while in the long-run the former are more reactive in terms of spatial mobility. The second paper assesses the impact of the crisis on the Eurozone labour markets integration by conducting a hierarchical cluster analysis. They observe that the last crisis has led to a polarisation of the Eurozone labour markets. Finally, the last study designs a classification approach based on a combination of regression and cluster analysis in order to identify idiosyncratic labour clusters to the Federal Employment Agency. In their two-step methodology, the greater the influence of an exogenous variable on the response variable in the regression analysis, the higher is the weight given to this variable in the cluster analysis. Within all this literature, our work can be inserted into the group of studies that, using labour matching data, generates labour clusters which can be useful for policy-making design and for the management of public employment agencies.

III. PROPOSAL

A. DATASETS

The data used for this purpose comes from the Continuous Working Life Sample (MCVL)[42], a large database containing micro-data on job matches which is provided by the Spanish Ministry of Employment, Migration and Social Security. The MCVL offers information from three Spanish public bodies: labour information from the Social Security system, administrative and personal information from the Continuous Municipal Register of Inhabitants and tax information from the National Tax Agency. The sample is published once a year and the population of reference is composed of individuals who have been paying contributions (such as registered workers or recipients of unemployment benefits) or receiving a contributory pension from Social Security at some date in the year of reference, regardless of how long they have been in that situation. The sample (in each year) comprises 4% of the people belonging to the reference population and is representative of the population registered at the Social Security system in the year of reference. The size of the sample exceeds one million people each year.

In this work, we use the MCVL information to know the characteristics of the workers and the jobs in the job placements that are registered in the Social Security system within the calendar year. The starting point in the processing of the MCVL data is to divide the workers and the jobs involved in the job matches into highly detailed groups according to their characteristics; groups which we call worker categories and job categories, respectively. Ideally, the detailed segmentation should allow us to consider the categories obtained as homogeneous or almost homogeneous, and the large size of the database should enable data (job matches) in each category to be sufficiently numerous as to be statistically representative.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

Therefore, our unit of analysis (which will be subject to clustering) is not going to be the individual worker (or the individual vacancy) but its category of belonging. When a job placement occurs, a match is generated between the worker's category and the job's category, a match that may imply a certain degree of occupational or geographical mobility. The availability of appropriate information on geographical and occupational labour mobility is an important requirement for the effectiveness of the labour matching process, and a prominent part of the active labour market policies (ALMPs).

After generating the categories of workers and jobs, the dataset is cross-classified in a contingency table where the rows represent worker categories (WC) and the columns represent job categories (JC). The cells of the contingency table are the frequencies (job matches) between the different categories of workers and jobs; i.e., the cell n_{ij} contains the number of job placements between the worker category w_i and the job category j_j.

As mentioned above, we have applied clustering techniques to two different periods, having each period its own dataset: 2011-2013, which corresponds to a period of economic crisis in Spain; and 2014-2016, which are years of economic recovery. The dataset of the period 2011-2013 contains 5,800 worker categories and 5,198 job categories with a total of 1,967,523 job placements. And the dataset of the period 2014-2016 is composed of 5,722 worker categories and 5,166 job categories with a total of 2,459,686 job placements.

B. MEHOTODOLOGY

The clustering analysis is carried out using two different clustering algorithms and two different technologies: k-means from Spark ML [43], and the average linkage algorithm included in Stata [44]. We have selected these two algorithms because, on the one hand, k-means is one of the most widely used partitioning algorithms, and the Spark version is implemented in a distributed manner and can be executed in a computer cluster. On the other hand, the average linkage is a hierarchical clustering method that has already been widely used in the literature [10, 11, 45, 46]. Therefore, they are widely contrasted clustering techniques and extensively used in many research fields.

These two algorithms have been executed taking the two datasets described in subsection III-A, so we have obtained two clustering results for each dataset. In order to analyse these clustering results, we have followed the methodology from [47]. The first step in a clustering process is to select the optimal number of clusters of each dataset. In the case of k-means from Spark, we have used two kinds of clustering validity indices (CVI), internal and external. We have applied the internal indices BD-Silhouette, BD-Dunn, Davies-Bouldin, and WSSSE included in [13]. In general terms, this kind of CVIs measures how the points are distributed through the clusters taking into account the compactness between the points and the separation between the clusters.

Let Ω be the space of the objects with a given distance d.

Then $\{A_k\}_{k=1..N}$ is a set of clusters so that $\bigcup_k A_k = \Omega$, and $A_i \cap A_j = \emptyset \quad \forall i \neq j$.

C_k is the centroid of A_k, and C_0 the centroid of Ω.

Let x_i be an element of A_k, $x_i \in A_k$, and let r_k be the distance from x_i to its own cluster A_k. Then, we can define the following CVIs:

- **BD-Silhouette** (BDS) (Eq 2): This index has been defined, for each possible partition, as the ratio between the difference of the $inter\text{-}cluster$ and the $intra\text{-}cluster$ distance, and the maximum of them.

$$BDS = \frac{inter\text{-}cluster - intra\text{-}cluster}{max\{inter\text{-}cluster, intra\text{-}cluster\}} \tag{2}$$

  where $inter\text{-}cluster$ (Eq 3) is the average of distances between each cluster centroid and the global centroid C_0:

$$inter\text{-}cluster = \frac{1}{N}\sum_{k=1}^{N} d(C_k, C_0) \tag{3}$$

  and $intra\text{-}cluster$ (Eq 4) distance is defined as the average of the distances of each point to the centroid of the cluster to which it belongs (Eq 5):

$$intra\text{-}cluster = \frac{1}{|N|}\sum_{x_i \in A_k}^{N} r_k \tag{4}$$

  where

$$r_k = \frac{1}{|A_k|}\sum_{x_i \in A_k} d(x_i, C_k) \tag{5}$$

  BD-Silhouette indicates an optimal value for the number of clusters on the first maximum, which maximises the coherence of the cluster with the lowest possible k.
- **BD-Dunn** (BDD): this index is given, for each possible partition, by the ratio between the minimum of the distances from the centroids to the global centre and the maximum of the distances from each point in the set to its centroid.

$$BDD = \frac{\min\limits_{k=1..N}\{d(C_k, C_0)\}}{\max\limits_{k=1..N}\max\limits_{x_i \in A_k}\{d(x_i, C_k)\}} \tag{6}$$

  BD-Dunn points out the number of clusters by the first maximum of the values.
- **Davies-Bouldin**(DB) [48]: In this index, we choose the first minimum of the Davies-Bouldin value chart to create a better model. The index is defined as follows:

$$DB = \frac{1}{N}\sum_{i}^{N}\sum_{j}^{N}\max_{i \neq j}\frac{r_i + r_j}{d(C_i, C_j)} \tag{7}$$

  where r_i and r_j are represented in Eq.5, and $d(C_i, C_j)$ is the distance between the centroids C_i and C_j.
- **Within Set Sum of Square Errors (WSSSE)** [43]: This index from Spark ML measures the cohesiveness of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

clusters and calculates the sum of the distances from each point to the centroid of its cluster. The optimal k is generally given by a global minimum or by the result after applying the elbow method to the WSSSE graph.

$$WSSSE = \sum_{x_i \in A_k} d(x_i, C_k)^2 \tag{8}$$

In addition, we have applied the external validity Chi-index to the k-means cluster. This kind of index measures how the points have been distributed by the clusters according to a given class variable. As for the average linkage clustering method, given its hierarchical nature, we have followed an internal validation method to select the optimal number of clusters.

Secondly, we have analysed the clustering results, taking into account the number of elements of each cluster and applying a descriptive statistical analysis. The third step is to evaluate the clustering results based on the features of the points of the datasets in both periods. In our case, we have considered the following features of the worker: region of residence (autonomous community and province), occupation group and sector of activity. Lastly, we have made a comparison between the clustering results for the k-means and the average linkage methods in the two periods.

IV. RESULTS

We have applied k-means from Apache Spark ML [43], and average linkage from [44]. This section includes the results obtained by following the methodology described above. This section is divided as follows: Subsection IV-A includes the clustering analysis using the k-means technique and shows the results for the sub-periods 2011-2013 and 2014-2016. Subsection IV-B follows the same structure of the previous subsection but the results are those of the average linkage method. Each of these subsections includes the selection of the optimal number of clusters, the description of the clustering results, and a comparison between the results of both sub-periods. Finally, Subsection IV-C carries out a comparison between the results of the k-means clustering and those coming from the average linkage clustering.

A. K-MEANS

Figure 3 shows the results of the internal CVIs of the k-means cluster for the sub-period 2011-2013. Each index is interpreted differently: Silhouette follows the "elbow method", which establishes the optimal number of clusters when the curve of the index begins to stabilise. In this case, Silhouette does not stabilise at any point until $k = 50$. The Dunn index points out the optimal number of clusters with local minimums; in this case, we can observe some local minimum along the curve, but we may not conclude that they are proper solutions because they are not decisive enough. Davies-Bouldin index points out the optimal number of clusters with local maximum, and as happened with the Dunn value, there are some local maximum but they do not look like suitable solutions because there are not determinant numbers. Finally, the WSSSE function points out the optimal solution as Silhouette, but the other way around, and we cannot find any stabilisation until $k = 50$. As can be observed, none of the indices concludes with an optimal number of clusters -we have found a similar situation for the next sub-period (2014-2016), so we have omitted the inclusion of the corresponding figure- so external CVIs need to be applied in order to find a proper clustering solution.

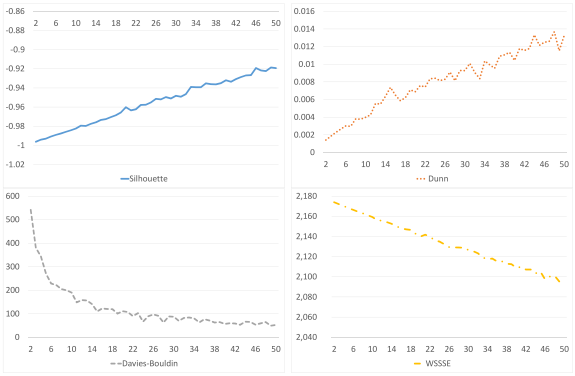


FIGURE 3: Internal CVI of k-means in the period 2011-13. X-axis represents the number of clusters and Y-axis the value of the index.

Figure 4 shows the results of the Chi Index [49] for the sub-period 2011-2013. Chi Index is defined as an external CVI which measures the quality of a clustering by means of the distribution of the instances through the classes, and the classes through the clusters. Chi Index measures the coherence between a class variable and a cluster through a contingency matrix. This matrix denotes the number of elements (job matches in our case) of each cluster (rows) in each value or category of the class variable (columns) in such a way that each cell ij of the matrix shows the total of matches of the cluster i in the category j of the class variable. Chi Index measures the coherence of this matrix dividing it into two components: the first one is a contingency matrix with relative values with respect to the marginal distribution of the clusters (represented by the blue line); the second one is the contingency matrix with relative values taking the marginal distribution of the class variable as reference (represented by the orange line). In this way, Chi Index is represented by two curves, which are the ones shown in the corresponding graphs of Figure 4. Chi Index was calculated assuming as classes: the region (Spanish Autonomous Communities), the province, the occupation group and the activity sector of the worker. Chi Index points out the optimal number of clusters (k) in the intersection between the curves.

We have decided not to show the graphs of the external index for the sub-period 2014-2016 because of space limitations; however, the results of this sub-period are included in Table 1. Table 1 represents the results of the Chi Index by each class in each sub-period. We have selected as the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

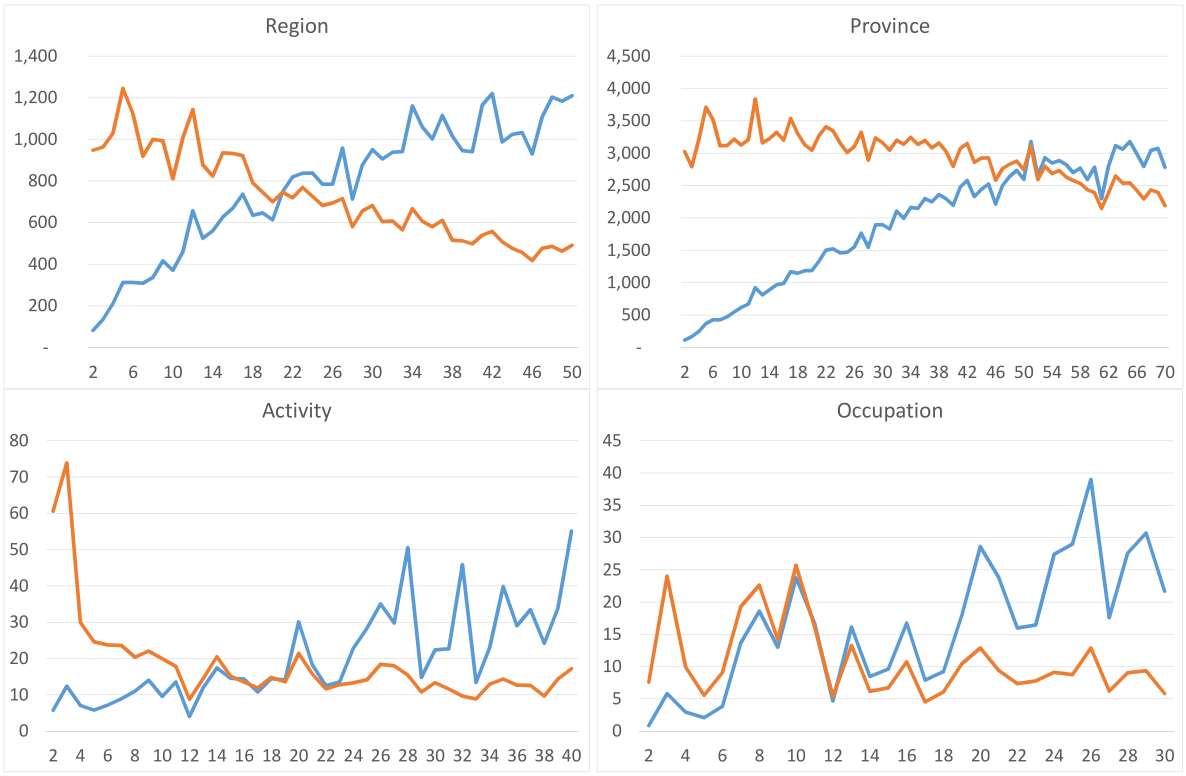


FIGURE 4: External clustering validity indices for k-means in the period 2011-13. X-axis represents the number of clusters. Y-axis shows the value of the index.

optimal number of clusters the one given by the region (rejecting province, occupation and sector classes) because it includes the province by definition, so that the province is directly located inside the region. In addition, the number of clusters given by the province (51 and 53 in each sub-period) is too large for having an easy to read and handle cluster solution. On the other hand, the activity and the occupation obtained lower numbers of clusters than the region, so, we may assume that these solutions are also included in the optimal number of clusters given by the region. Hence, we have considered $k = 21$ for the sub-period 2011-2013, and $k = 22$ for the next sub-period, 2014-2016.

Class	2011-2013	2014-2016
Region	**21**	**22**
Province	51	53
Activity	16	13
Occupation	11	10

TABLE 1: Results of the external validity clustering indices for the periods 2011-13 and 2014-16. In bold, the chosen result.

Table 2 shows the number of worker categories and job placements for the clusters $k = 21$ means in the sub-period 2011-2013, and $k = 22$ means in the sub-period 2014-2016. It is worth mentioning that the clusters with the same identification number in both sub-periods are not the same, the number is just used to name them. In addition, it must be observed that the sub-period 2014-2016 has got one cluster more than the previous sub-period.

Focusing on the sub-period 2011-2013, the 5,800 WCs have been homogeneously distributed across all the clusters. Clusters have an average size of 276 WCs. Cluster 14 is the smallest one with 122 WCs (2% of the total), and cluster 1 has got the highest number of elements, with 446 WCs (11%). In general terms, the job placements are in line with the size of the cluster, with the largest group being the one with the largest number of job placements. On the other hand, the result for $k = 22$ in the sub-period 2014-2016 does not differ very much from the previous scenario. As can be observed, the clusters of this sub-period are composed of 260 WCs on average, with a range between 114 WCs (cluster 9) and 602 WCs (cluster 19), and between 2,226 matches (cluster 2) and 388,434 matches (cluster 19).

A summary of the cluster structure by region, province, activity sector and occupation group can be found in the Tables 3 and 4, one for each sub-period. These tables are built by considering for each cluster only those categories of the variables with the highest percentages in terms of job matches. Specifically, they show the id number of the cluster; the size of the cluster in terms of job matches, which was set in intervals of the equal width, starting with the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

	2011-2013		2014-2016	
#	WCs	Placements	WCs	Placements
0	331	74,545	229	56,486
1	**446**	174,900	350	213,003
2	176	73,095	225	**2,226**
3	391	237,447	173	33,686
4	414	197,624	142	57,341
5	132	51,316	195	61,578
6	430	43,664	340	67,956
7	256	39,617	205	74,264
8	306	99,857	308	118,065
9	342	56,912	**114**	8,624
10	207	80,621	247	111,389
11	267	48,643	289	152,245
12	341	159,071	196	85,523
13	124	30,808	343	199,403
14	**122**	31,493	183	28,121
15	213	71,198	244	65,130
16	141	**8,024**	379	147,638
17	250	132,465	243	173,358
18	385	**268,371**	400	350,751
19	298	63,395	**602**	**388,434**
20	228	24,457	133	30,246
21	-	-	182	34,219
Total	5,800	1,967,523	5,722	2,459,686
Avg	276.19	93,692	260.09	111,803.91
Min	122	8,024	114	2,226
Max	446	268,371	602	388,434

TABLE 2: Clustering results for k-means with k=21 in the period 2011-13 and with k=22 in the period 2014-16. Minimum and maximum of each column are highlighted in bold.

size of the smallest cluster (114), so that, 'S' is set for small clusters in the range $[114, 228]$, medium (M) within the range $(229, 343]$, and large (L), for those clusters larger than 343 elements; the main locations of the cluster in cardinal points form; the ids of the main sectors of activity, whose respective assignment can be found in Table 12; and the main occupation groups of the clusters which have been grouped in the following categories: Managers and workers with university degree (UnivDegr), Technical engineers and qualified assistants (TechEngin), Clerical and workshop heads (C&WHeads), and the rest of occupations, which have been categorised as Low-skilled.

Table 3 shows the clustering features for the sub-period 2011-2013. As can be seen, there exist five large clusters (1, 3, 4, 6 and 18) geographically distributed in different spatial areas of the country. It is interesting to note that these clusters are mainly composed of low-skilled workers, with the exception of clusters 6 and 18 which add Technical Engineers and C&WHeads respectively. In addition, there is no predominant sector of activity, although the most common sector is the manufacturing industry, which is present in 3 of these 5 clusters. It should be highlighted that clusters 5 and 10 are mainly based on workers from the Canary Islands and the Balearic Islands respectively. Besides that, the clusters 0, 6, 16 and 20 are the only ones with Technical Engineers, and just the clusters 0 and 20, which are located in the Centre of Spain, are in addition composed of University degrees. It is noteworthy that agricultural workers are mainly located in the Centre and the South of the country, as the clusters 2, 7, and 17 show. It is also interesting to point out that there are four clusters (0, 14, 16, and 18) whose principal occupation group is C&WHeads; they mainly share the Financial & Business Services activity and do not have a predominant location.

Table 4 summarises the features of the clustering for the sub-period 2014-2016. In general, the clusters are from just one location, and when there is more than one location, they show geographical proximity. The largest clusters are mainly composed of worker categories from the North but the cluster 19 that is composed of Southern (including Ceuta) and Balearic workers. All these clusters have a non-qualified occupation group as predominant, except the cluster 6 that is just of higher education. In this period, there are more small clusters than medium ones, and the cluster size is related neither to the occupation group nor to the sector of activity. In this sub-period, there are also three clusters (4, 6, and 9) whose occupation group is mainly composed of workers with university studies; the main location of these clusters is the North of the country and they do not share any specific economic activity.

1) 2014-16 vs 2011-13

This section carries out the comparison between the clusters of the sub-periods 2011-2013 and 2014-2016. Table 5 shows the clusters from the sub-period 2014-2016 and their correspondence with the clusters of the previous sub-period in terms of the worker categories that they have in common. The colour of the cells indicates the level of relationship between the clusters, the darker the green, the stronger the relationship-the greater is the number of the worker categories that they have in common.

It should be highlighted that there are six clusters (3, 4, 10, 15, 17, and 20) which have correspondence only with one of the clusters of the previous sub-period, although this correspondence is not 100% or one-to-one, since the corresponding clusters of the first sub-period (2011-2013) with which the six clusters match also appear related to other clusters of the second sub-period analysed (2014-2016); in other words, some clusters of sub-period 2011-2013 have been separated into different clusters of the sub-period 2014-2016. For instance, the cluster 17 of the second sub-period (2014-2016), which is mainly composed of low-skilled individuals working at construction and industrial manufacturing in the East side of the country, belongs to a larger cluster (the number 12) of the first sub-period (2011-2013) which also keeps some correspondence with the cluster 11 (2014-2016); cluster 12 (2011-2013) is a cluster of low-skilled workers from the industrial manufacturing sector. We find a similar result with cluster 15, which belongs to cluster 9 of the sub-period 2011-2013; in both clusters we find low-skilled workers in the central area and working in the educational sector.

On the other hand, the clusters 1, 6, 8, and 9 of sub-period 2014-2016 are linked with multiple clusters of the first sub-period (2011-2013). This result may indicate that larger local labour markets have emerged in this second sub-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

#	Size	Location	Activity	Occupation
0	M	Centre	Serv3 / Serv2	TechEngin / UnivDegr / C&WHeads
1	L	North-East	PublicAd / Serv1 / Constr / Manuf	Low-skilled
2	S	South	Agric	Low-skilled
3	L	South / Ceuta	Constr / PublicAd / Serv3	Low-skilled
4	L	Centre	PublicAd / Educ / Health	Low-skilled
5	S	Balearic I.	Supplies / PublicAd / Serv1	Low-skilled
6	L	North / South / Centre	Manuf / Supplies	TechEngin / Low-skilled
7	M	Centre	Agric / Health / Serv1	Low-skilled
8	M	Northeast / Northwest	PublicAd / Educ / Serv3	Low-skilled
9	M	Centre	PublicAd / Constr / Educ / Serv3	Low-skilled
10	S	Canary I.	PublicAd / Serv1 / Educ / Serv3	Low-skilled
11	M	Centre / East	Serv1	Low-skilled
12	M	South / East	Manuf	Low-skilled
13	S	Centre	Constr	Low-skilled
14	S	North	Serv2	C&WHeads
15	S	North	Educ / Manuf / Serv3 / Health	Low-skilled
16	S	Centre / South	Serv1 / Manuf / Serv2	TechEngin / C&WHeads
17	M	Centre / South	Agric	Low-skilled
18	L	Northeast	Manuf	C&WHeads / Low-skilled
19	M	Northwest / Centre	PublicAd / Serv1 / Constr	Low-skilled
20	S	Centre	PublicAd / Educ	TechEngin / UnivDegr

TABLE 3: Summary of the clustering features of k-means with $k = 21$ in the period 2011-13.

#	Size	Location	Activity	Occupation
0	M	North / Centre	Agric	Low-skilled
1	L	East / Canary I.	Constr	Low-skilled
2	S	Northwest / Centre	Supplies / ExtratOrg	Low-skilled / UnivDegr / C&WHeads
3	S	Centre	Serv1 / PublicAd	Low-skilled
4	S	South	Educ / Health	TechEngin / UnivDegr
5	S	North	Educ / Health / Manuf	Low-skilled
6	L	North / Centre	Serv3	TechEngin / UnivDegr
7	S	Centre	Agric	Low-skilled
8	M	North	Manuf / Serv2	Low-skilled
9	S	Northeast	Serv2	TechEngin / UnivDegr / C&WHeads
10	M	Centre	Constr	Low-skilled
11	M	South	Agric / Educ / Serv1	Low-skilled
12	S	Centre / Canary I.	Serv1 / Constr	Low-skilled
13	M	Centre	Serv1 / Manuf / Serv3 / Serv2	TechEngin / C&WHeads / Low-skilled
14	S	Northeast	PublicAd / Serv1	Low-skilled
15	M	Centre	Educ	Low-skilled
16	L	Northwest	PublicAd	Low-skilled
17	M	East	Construc / Manuf	Low-skilled
18	L	Northwest	PublicAd / Serv1 / Serv3	Low-skilled
19	L	South / Ceuta / Balearic I.	PublicAd / Constr / Serv1	Low-skilled
20	S	Centre	Health	TechEngin / C&WHeads
21	S	Centre	Health / Serv1	Low-skilled

TABLE 4: Summary of the clustering features of k-means with $k = 22$ in the period 2014-16.

period. For example, the cluster 8 from sub-period 2014-2016, which is mainly formed by low-skilled workers from the Northern area in the industrial manufacturing, and financial and business services, is composed of worker categories of the clusters 1, 14 and 15 of sub-period 2011-2013, which are located in the North, with low-skilled workers in most cases and with a range of different economic activities. In the case of the cluster 9, which is composed of high education workers from the Northeast in the activity of financial and business services, is formed of clusters 0, 18 and 9 of the first sub-period; these clusters mainly have workers with higher education, are also dedicated to the financial and business services and share some geographical locations.

The rest of the clusters (0, 5, 7, 12, 13, 14, 16, 18, 19 and 21) are related to two clusters of the first sub-period, although only with one of them maintain a strong relationship. For instance, cluster 5, which is composed of low-skilled workers from the Northern area and is dedicated to education, health and industrial manufacturing, is formed by clusters 1 and 15 of the sub-period 2011-13, which are clusters from the North, and share similar economic activities and occupation groups. We find a similar situation with cluster 16, which is from the Northwest, dedicated to public administration, and it is composed of clusters 8 and 19 of the sub-period 2011-13, which are also from the Northwest and belong to the public administration sector. Another case is the one of cluster 19 from 2014-16 (large size), which is mainly composed of cluster 3 and, to a lesser extent, of cluster 5. These clusters are located in the South, Ceuta and Balearic Islands, and belong to the public administration and construction sectors.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

2014-16	2011-13		
0	6	19	
1	1	10	11
2	6		
3	11		
4	2		
5	1	15	
6	7	20	1
7	2	7	
8	15	14	1
9	0	18	9
10	4		
11	17	12	6
12	10	0	
13	4	0	
14	18	8	
15	9		
16	8	19	
17	12		
18	18	8	
19	3	5	
20	20		
21	13	7	

TABLE 5: Correspondence between the k-means clusters of the periods 2014-16 and 2011-13.

We can conclude that due to the change in the cycle of the economy, there have been some movements in the Spanish labour market which have changed the physiognomy of some of the 'job creation' clusters. However, there still exist some clusters that remain stable despite the economic crisis (showing some degree of inertia).

B. AVERAGE LINKAGE

This section follows a similar structure than the previous one. Firstly, it contains the study of the optimal number of clusters for both sub-periods, and then, the optimal clustering result is described.

Figure 5 relates the inter-cluster and the intra-cluster distances for each possible number of clusters. The blue line in the figure represents the ratio between both distances, and the red line represents the increase of that ratio in percent. In this case, we have chosen $k = 191$ as the optimal number of clusters because a proper solution is given by a highest ratio between the inter-cluster and the intra-cluster distances until its increment stop raising, so that the red line tends to zero. After the selection of $k = 191$, we have only taken the 23 clusters that have 1,500 or more job placements. In this way, we keep 99% of the job placements, and just skip those clusters which contain few elements. We must bear in mind that choosing the 23 largest clusters for $k = 191$ is not the same as initially estimating $k = 23$.

In the same way, Figure 6 shows the results for the period 2014-16, where we have taken $k = 176$ as the optimal number of clusters; of those clusters, we have analysed the 25 largest-those with more than 1,500 job placements-.

Table 6 shows the results for the average linkage with $k = 23$ and $k = 25$ in the sub-periods 2011-13 and 2014-16, respectively. The data analysed with the average linkage

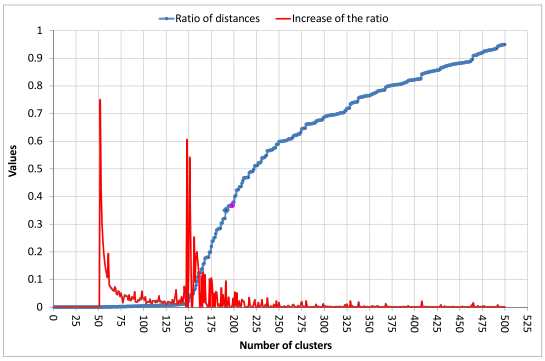


FIGURE 5: Representation of the selection of the optimal number of clusters in the period 2011-13. The blue line represents the ratio between inter-cluster and the intra-cluster distances for each possible number of clusters, and the red line represents the increase of that ratio in percent. The chosen optimal number of clusters was 191 of which we have studied 23.

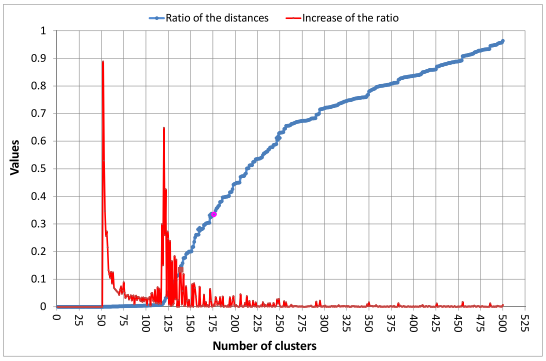


FIGURE 6: Representation of the selection of the optimal number of clusters in the period 2014-16. The blue line represents the ratio between inter-cluster and the intra-cluster distances for each possible number of clusters, and the red line represents the increase of that ratio in percent. The chosen optimal number of clusters was 176 of which we have studied 25.

method for the first sub-period is composed of a total of 5,317 worker categories that give rise to 1,941,816 job matches. The 23 clusters have got 231 WCs and more than 84,000 job placements on average. The cluster 41 is the one with the fewest number of WCs, just 42, and cluster 11 is the one with the fewest number of matches (4,940). As can be observed, the clusters with more WCs and job placements do not match either: cluster 1, which has the largest number of WCs, contains 801 WCs with 289,446 job placements, while cluster 12, the one with the largest number of matches, is composed of 531 WCs with 304,996 job placements.

On the other hand, 5,486 worker categories are analysed during the sub-period 2014-16. In this case, the clusters have 219 WCs and 98,257 job placements on average. Cluster 7

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

is the one with more WCs and placements, 878 and 388,935 respectively. Moreover, the clusters with the lowest number of WCs and placements do not match: cluster 4 contains only 3 WCs and 1,815 matches, and cluster 24 has 43 WCs and just 1,640 matches.

2011-2013			2014-2016		
#	WCs	Placements	#	WCs	Placements
1	**801**	289,446	1	296	120,188
2	246	16,683	2	87	28,378
3	75	30,405	3	221	60,089
4	186	47,340	4	**3**	1,815
6	189	55,131	7	**878**	**388,935**
8	86	10,855	9	192	70,966
9	69	5,543	10	23	2,328
10	80	20,123	11	42	18,352
11	67	**4,940**	13	242	42,833
12	531	**304,996**	14	133	18,880
14	307	119,781	15	115	41,140
16	198	90,041	16	110	17,666
19	489	298,560	17	381	201,040
20	108	49,342	18	228	160,443
21	289	53,035	19	214	103,842
22	210	123,536	21	312	233,901
23	222	110,822	22	521	282,188
24	164	26,685	24	43	**1,640**
25	262	94,269	25	492	375,928
27	93	13,007	26	122	62,932
29	205	46,134	28	270	65,786
34	398	112,161	32	385	126,576
41	**42**	18,981	34	64	3,407
-			38	48	23,699
-			44	64	3,482
Total	5,317	1,941,816	Total	5,486	2,456,434
Avg	231.17	84,426.78	Avg	219.44	98,257.36
Min	42	4,940	Min	3	1,640
Max	801	304,996	Max	878	388,935

TABLE 6: Clustering result for average linkage with $k = 23$ and $k = 25$ in the periods 2011-13 and 2014-16, respectively. Minimum and maximum of each column are highlighted in bold.

Table 7 summarises the features of the clustering result for the sub-period 2011-13 with $k = 23$. In this case, the size of the clusters is evenly divided. There is just one large cluster, 7 medium clusters and 15 small clusters. It should be highlighted that there are 11 clusters which are composed of worker categories from the centre of Spain. There are just 4 clusters (2, 3, 8, and 41) with university studies as principal occupation group, of which three of them are located in the Centre and the South of the country, and their main activities are health, manufacturing and some services. Likewise, there are two clusters (11 and 20) whose main occupation group is C&Wheads (the main activity is trade, transport, accommodation and communication), but one is placed in the Centre and the other in the Balearic Islands. The rest of the clusters (17 clusters) have no high education levels among their main occupations: three of them (6, 12, and 14) are based on agriculture (West or South location and medium or small-size); other one (small) is composed of Canary workers in the sector of trade, transport, accommodation, communication, and other services; and two of them (22 and 23) are from the East and share the industrial manufacturing sector as main economic activity, among others.

Next, the features of the clustering for the sub-period 2014-16 are going to be discussed (Table 8). In this case, we find 3 large clusters, 5 medium clusters, and 17 small clusters. The clusters 4, 10, 11 and 38 are the only ones with university studies; in addition, their main sector of activity is health, and they are located all around Spain. The other clusters do not have, in general, high level of studies. In these clustering results, we find several clusters located only in one province, such as the 19 (Canary Islands), the 26 (Balearic Islands), the 24 (Ceuta), and the cluster 44 (Melilla); all of them are mainly focused on the sector of trade, transport, accommodation and communication. The average linkage clustering in this sub-period also includes two clusters (21 and 22) from the Southern zone whose principal economic activity is agriculture, although they also include workers of the sectors of construction and education.

C. K-MEANS VS AL CLUSTERS

This section includes a comparison between the results of the k-means (KM) and the average linkage (AL) methods in both sub-periods. Tables 9 and 10 show the correspondence between the results of the k-means and the average linkage clusters during the periods 2011-13 and 2014-16 respectively. As mentioned above, the colour of the cells indicates the level of relationship between those clusters, the darker the green, the stronger the relationship.

Table 9 shows the comparison for the sub-period 2011-2013. The KM clusters 3, 4, 5, 9, 10, 13, 14, 15, 16, and 18 are directly related with just one AL cluster. This indicates that the clustering results of the KM are similar to those of the average linkage. There are 6 KM clusters that are composed of two AL clusters, but only with one of them, the relationship can be considered strong. Just the KM clusters 0, 6, and 20 have got a weak relationship with the AL cluster.

We find a similar situation in the second sub-period (2014-2016), which is represented in Table 10. The KM clusters 3, 4, 7, 10, 11, 13, 16, 17 and 18 are directly related with just one AL cluster. Likewise, the KM clusters 0, 1, 5, 8, 12, 14, 15, 20, and 21 are composed of two AL clusters; i.e., the worker categories of some AL clusters are joined to build a new KM cluster. Finally, there are just 3 KM clusters (2, 9 and 19) that do not have a strong relationship with any specific AL cluster; they match with several AL clusters but at a very low rate.

In order to quantify the similarity between the clustering solutions (KM and AL) of each sub-period, we have calculated the ratio of coincidence between those solutions. For that purpose, we have considered the similarity between a KM cluster and an AL cluster as the ratio of the elements (worker categories) in common in relation to the total of elements of the KM cluster. This comparison can also be done on a scale of one-to-many (one KM cluster and several AL units), so, for each KM cluster, we have progressively taken from 1 to 3 AL clusters (sorted from the highest to the lowest relation with the KM cluster) in order to calculate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

#	Size	Location	Activity	Occupation
1	L	Centre	Serv3 / Educ	Low-skilled
2	M	Centre / South	Serv1 / Manuf / Serv2	UnivDegr / TechEngin
3	S	Centre / South	Health	TechEngin
4	S	North / Centre	Health / Constr / Serv3	Low-skilled
6	S	West	Agric / Educ / PublicAd	Low-skilled
8	S	Centre	Manuf / Serv1 / Serv3	TechEngin
9	S	Centre	Serv1 / Health / Constr	Low-skilled
10	S	North / Centre	Manuf / Constr	Low-skilled
11	S	Centre	Serv1 / PublicAd	C&WHeads
12	M	South	PublicAd / Agric / Serv3	Low-skilled
14	M	South	Agric / Serv1	Low-skilled
16	S	Canary I.	Serv1 / Serv3	Low-skilled
19	M	Northeast	PublicAd	Low-skilled
20	S	Balearic I.	Serv1 / Serv3	C&WHeads
21	M	Northeast	PublicAd / Serv3	Low-skilled
22	S	East	Manuf / Serv3	Low-skilled
23	S	East	Constr / Manuf	Low-skilled
24	S	Centre	Serv1	Low-skilled
25	M	North	Educ / Serv3 / Health	Low-skilled
27	S	Centre	Health / Serv1 / Educ	Low-skilled
29	S	North	Constr / Manuf	Low-skilled
34	M	Northwest	Constr	Low-skilled
41	S	Centre / South	Health	UnivDegr

TABLE 7: Summary of the clustering features of average linkage with $k = 23$ in the period 2011-13.

#	Size	Location	Activity	Occupation
1	M	North	Health / Serv1	Low-skilled
2	S	North	Manuf / Constr	Low-skilled
3	S	North	Agric	Low-skilled
4	S	North	Health / Constr / Serv3	TechEngin / C&Wheads
7	L	Centre	Constr / Educ / Serv3	Low-skilled
9	S	West	Serv1 / PublicAd	Low-skilled
10	S	Centre / South	PublicAd	UnivDegr
11	S	Centre / South	Health	TechEngin / UnivDegr
13	S	Centre	Serv1	Low-skilled
14	S	Centre	Serv1 / PublicAd	Low-skilled
15	S	North	Serv2 / Constr	Low-skilled
16	S	Centre	Health / Manuf / Serv1	Low-skilled
17	M	Centre / East	Manuf / Serv1	Low-skilled
18	S	East	Educ	Low-skilled
19	S	Canary I.	Serv1	Low-skilled
21	M	South	Constr / Agric	Low-skilled
22	L	South	Educ / Agric	Low-skilled
24	S	Ceuta	PublicAd / Serv1	Low-skilled
25	L	Northeast	PublicAd	Low-skilled
26	S	Balearic I.	Serv1 / Educ / Manuf	Low-skilled
28	M	Northeast	PublicAd / Serv1	Low-skilled
32	M	Northwest	Serv1 / PublicAd	Low-skilled
34	S	Centre	Agric	Low-skilled
38	S	All	Health	UnivDegr
44	S	Melilla	Serv1 / PublicAd	Low-skilled

TABLE 8: Summary of the clustering features of average linkage with $k = 25$ in the period 2014-16.

the corresponding ratios. Table 11 shows the results of our comparison. The different sub-periods are represented by rows, and the number of clusters that we have taken to make the comparison is expressed by columns. In the period 2011-2013, we have obtained a 66% of similarity between the KM clusters and AL clusters taking just the AL unit which has the highest number of common worker categories. Taking 2 AL clusters, we have obtained that the clusters are similar by 83%. Finally, taking 3 AL clusters, we have obtained a similarity rate of 90%. Furthermore, we find a similar picture for the period 2014-2016, but with higher percentages, ranging from 71% of similarity if we take the AL cluster with the highest rate to 98% if we take 3 AL clusters.

V. CONCLUSIONS

In this study, a labour matching analysis of the Spanish labour market is developed based on the recent labour matching flow. This analysis may allow the authorities to orientate, geographically and occupationally, the worker's search. We have applied an unsupervised machine learning technique, such as the clustering methodology, with the aim to discover how the labour market is organised, taking as unit of analysis the different categories of the workers who get a job. The ini-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

K-means	Average Linkage		
0	1	2	8
1	29	22	10
2	6	12	
3	12		
4	1		
5	20		
6	4	12	14
7	1	6	9
8	34	19	
9	21		
10	16		
11	24	22	
12	23	14	
13	1		
14	25		
15	25		
16	2		
17	14	12	
18	19		
19	34	4	
20	27	11	1

TABLE 9: Correspondence between the 21 clusters of k-means and 23 clusters of AL in the period 2011-13.

K-means	Average Linkage		
0	15	16	
1	18	19	
2	32	7	
3	17		
4	21		
5	3	1	
6	13	7	2
7	9		
8	1	3	
9	25	7	3
10	7		
11	22		
12	19	14	
13	7		
14	25	28	
15	28	34	
16	32		
17	17		
18	25		
19	22	21	26
20	13	11	
21	7	14	

TABLE 10: Correspondence between the 22 clusters of k-means and 25 clusters of AL in the period 2014-16.

Similarity	1 Cluster	2 Clusters	3 Clusters
2011-2013	66%	83%	90%
2014-2016	71%	91%	98%

TABLE 11: Similarity between the clustering results of the k-means and the average linkage.

Activity sector	Id
Agriculture	Agric
Construction	Constr
Education	Educ
Extraterritorial Organisations	ExtratOrg
Health	Health
Manufacturing	Manuf
Mining	Mining
Public Administration	PublicAd
Trade, Transport, Accommodation & Communication	Serv1
Financial & Business Services	Serv2
Other Services	Serv3
Supplies	Supplies

TABLE 12: List of sectors of activity with their assigned code.

tial databases have been pre-processed to work with worker and job categories which are related through a contingency table that contains the job placements that occur between them, representing a two-sided matching model. We have applied two different clustering algorithms, with different technologies. Thereby, with each clustering algorithm, we have applied different methods to discover the optimal number of clusters. Then, we have characterised the clustering results, focusing on the size of the clusters, the geographical location, the activity sector, and the occupation group of the workers. Finally, we have made a comparison between the different periods to see the evolution of the labour market under both clustering methods. Our methodology is versatile and could be adapted to many other labour analyses.

The findings of this study provide evidence of the effects of the recent economic crisis in the Spanish labour market. One could conclude from these results that there have been some transformations in the Spanish labour market, which have changed the physiognomy of some of the "job creation" clusters. However, there still exist some clusters that remain stable despite the economic crisis. These movements have been observed in the results of both clustering methods. In addition, there exists a strong similarity between the k-means and average-linkage results, in such a way that the ratio of similarity was between the 66% and 98% depending on the number of AL clusters that we take into account. These two approaches can support the economic and political decision making in different public administrations, as well as the customisation of the employment policies, improving the ALMPs.

Finally, we have also achieved an interesting characterisation of groups of workers all around Spain. In this sense, out methodology is also useful to capture the structure of the labour market (local labour markets, for instance).

ACKNOWLEDGMENTS

This work has been supported by the Spanish Ministry of Economy and Competitiveness under projects TIN2014-55894-C2-R and TIN2017-88209-C2-2-R. J.M. Luna-Romera holds a FPI scholarship from the Spanish Ministry of Economy and Competitiveness. Fernando Núñez-

Hernández and Carlos Usabiaga acknowledge the funding from the Spanish Ministry of Economics, Industry and Competitiveness (ECO2017-86780-R Project) and the Andalusian Government (SEJ-513 PAIDI Research Group).

REFERENCES

[1] European Union. Total unemployment rate. https://ec.europa.eu/eurostat, 2018. [Online; accessed 18-october-2018].

[2] CIS. Three principal problems. http://www.cis.es/, 2018. [Online; accessed 18-october-2018].

[3] Barbara Petrongolo and Christopher A. Pissarides. Looking into the black box: A survey of the matching function. Journal of Economic Literature, 39(2):390–431, 2001.

[4] Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[5] J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[6] I.H. Witten, E. Frank, M.A. Hall, and C.J. Pal. Data Mining: Practical Machine Learning Tools and Techniques. 2016.

[7] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2):165–193, Jun 2015.

[8] Nisha and Puneet Jai Kaur. A survey of clustering techniques and algorithms. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pages 304–307, 2015.

[9] Jiawei Han, Micheline Kamber, Jian Pei, Jiawei Han, Micheline Kamber, and Jian Pei. 10 – Cluster Analysis: Basic Concepts and Methods. In Data Mining, pages 443–495. Morgan Kaufmann, 2012.

[10] A. Triayudi and I. Fitri. Comparison of parameter-free agglomerative hierarchical clustering methods. ICIC Express Letters, 12(10):973–980, 2018.

[11] Benjamin Moseley and Joshua Wang. Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 3094–3103. Curran Associates, Inc., 2017.

[12] Junjie Wu, Jian Chen, Hui Xiong, and Ming Xie. External validation measures for k-means clustering: A data distribution perspective. Expert Systems with Applications, 36(3, Part 2):6050 – 6061, 2009.

[13] José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros, and José C. Riquelme Santos. An approach to validity indices for clustering techniques in big data. Progress in Artificial Intelligence, 7(2):81–94, 2018.

[14] Theresa J. Devine and Nicholas M. Kiefer. Empirical labor economics: The search approach. Oxford University Press, New York, 1991.

[15] Dale Mortensen and Christopher Pissarides. New developments in models of search in the labor market. In O. Ashenfelter and D. Card, editors, Handbook of Labor Economics, volume 3, Part B, pages 2567–2627. Elsevier, Amsterdam, 1 edition, 1999.

[16] Richard Rogerson, Robert Shimer, and Randall Wright. Search-theoretic models of the labor market: A survey. Journal of Economic Literature, 43(4):959–988, 2005.

[17] Eran Yashiv. U.S. labor market dynamics revisited. The Scandinavian Journal of Economics, 109(4):779–806, 2007.

[18] Christopher A. Pissarides. Equilibrium unemployment theory. MIT Press, Cambridge, Mass, 2000.

[19] Christopher Pissarides. EconomicDynamics Interviews Christopher Pissarides on the Matching Function. EconomicDynamics Newsletter, 10(1), 2008.

[20] Robert Shimer. Mismatch. American Economic Review, 97(4):1074–1101, 2007.

[21] Robert Lucas and Edward Prescott. Equilibrium search and unemployment. Journal of Economic Theory, 7(2):188–209, 1974.

[22] Dale T. Mortensen. Island matching. Journal of Economic Theory, 144(6):2336–2353, 2009.

[23] Ricardo Lagos. An alternative approach to search frictions. Journal of Political Economy, 108(5):851–873, 2000.

[24] P.A. Gautier. Non-sequential search, screening externalities and publications and the public good role of recruitment offices. Economic Modelling, 19(2):179–196, 2002.

[25] Michael Sattinger. Queueing and searching. pages University at Albany, SUNY, 2010.

[26] Melvyn Coles and Eric Smith. Marketplaces and matching. International Economic Review, 39(1):239–54, 1998.

[27] Ehsan Ebrahimy and Robert Shimer. Stock-flow matching. Journal of Economic Theory, 145(4):1325–1353, 2010.

[28] Pablo Álvarez de Toledo, Fernando Núñez, and Carlos Usabiaga. La función de emparejamiento en el mercado de trabajo español. Revista de Economia Aplicada, 16(3):5–35, 2008.

[29] Pablo Álvarez de Toledo, Fernando Núñez, and Carlos Usabiaga. An Empirical Analysis of the Matching Process in the Spanish Public Employment Agencies: The Vacancies. Working Papers 11.03, Universidad Pablo de Olavide, Department of Economics, Seville, 2011.

[30] R Cotterman and Franco Peracchi. Classification and aggregation: An application to industrial classification in cps data. Journal of Applied Econometrics, 7(1):31–

51, 1992.

[31] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. ACM Comput. Surv., 31(3):264–323, 1999.

[32] Pablo Álvarez de Toledo, Fernando Núñez, and Carlos Usabiaga. An empirical approach on labour segmentation. Applications with individual duration data. Economic Modelling, 36(C):252–267, 2014.

[33] Pablo Álvarez de Toledo, Fernando Núñez, and Carlos Usabiaga. Matching and clustering in square contingency tables. Who matches with whom in the Spanish labour market. Computational Statistics & Data Analysis, 127:135–159, 2018.

[34] Dominic Power and Mats Lundmark. Working through knowledge pools: Labour market dynamics, the transference of knowledge and ideas, and industrial clusters. Urban Studies, 41(5-6):1025–1044, 2004.

[35] Rikard Eriksson and Urban Lindgren. Localized mobility clusters: Impacts of labour market externalities on firm performance. Journal of Economic Geography, 9(1):33–53, 2009.

[36] Octávio Figueiredo, Paulo Guimarães, and Douglas Woodward. Firm–worker matching in industrial clusters. Journal of Economic Geography, 14(1):1–19, 2014.

[37] Peter J. Sloane, Philip D. Murphy, Ionnis Theodossiou, and Michael White. Labour market segmentation: A local labour market analysis using alternative approaches. Applied Economics, 25(5):569–581, 1993.

[38] A. Chakraborty, M. A. Beamonte, A. E. Gelfand, M. P. Alonso, P. Gargallo, and M. Salvador. Spatial interaction models with individual-level data for explaining labor flows and developing local labor markets. Comput. Stat. Data Anal., 58:292–307, 2013.

[39] Hector Sala and Pedro Trivín. Labour market dynamics in Spanish regions: Evaluating asymmetries in troublesome times. SERIEs, 5(2):197–221, 2014.

[40] Hélène Syed Zwick and S. Ali Shah Syed. The polarization impact of the crisis on the Eurozone labour markets: A hierarchical cluster analysis. Applied Economics Letters, 24(7):472–476, 2017.

[41] Uwe Blien and Franziska Hirschenauer. A new classification of regional labour markets in Germany. Letters in Spatial and Resource Sciences, 11(1):17–26, 2018.

[42] Ministerio de Trabajo, Migraciones y Seguridad Social. La Muestra Continua de Vidas Laborales. Estadísticas, Presupuestos y Estudios, Seguridad Social, Madrid, Spain, 2016.

[43] Apache Spark. Clustering - Spark 2.3.0 Documentation. https://spark.apache.org/docs/2.3.0/ml-clustering.html, 2018. [Online; accessed 12-july-2018].

[44] Clustering Stata. Cluster analysis - Stata. https://www.stata.com/features/cluster-analysis/, 2018. [Online; accessed 12-july-2018].

[45] P. Yildirim and D. Birant. K-linkage: A new agglomerative approach for hierarchical clustering. Advances in Electrical and Computer Engineering, 17(4):77–88, 2017.

[46] X. Wu, T. Ma, J. Cao, Y. Tian, and A. Alabdulkarim. A comparative study of clustering ensemble algorithms. Computers and Electrical Engineering, 68:603–615, 2018.

[47] Rubén Pérez-Chacón, José M. Luna-Romera, Alicia Troncoso, Francisco Martínez-Álvarez, and José C. Riquelme. Big data analytics for discovering electricity consumption patterns in smart cities. Energies, 11(3), 2018.

[48] D. L. Davies and D. W. Bouldin. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2):224–227, 1979.

[49] José María Luna-Romera, María Martínez-Ballesteros, Jorge García-Gutiérrez, and José C. Riquelme. External clustering validity index based on chi-squared statistical test. Information Sciences, 487:1 – 17, 2019.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2935386, IEEE Access

JOSÉ MARÍA LUNA-ROMERA is a PhD research student at University of Sevilla (Spain) since November 2015 after being awarded a four-year research scholarship by the Spanish Government. He received his M.Sc. degree in Software Engineering and Technology in 2012 and published his master dissertation on Data Mining Applied to Earthquakes Prediction. His current research interests concern clustering analysis, and more generally data mining and Big Data.

FERNANDO NÚÑEZ BA (1997) and PhD (2006, with distinction) in Economics. Economics lecturer since 2001. Research stays, among other places, at Carlos III University (Spain) and the Universities of Manchester and Essex (both in the UK). Current academic secretary of the Department of Industrial Organization and Business Management I at the University of Seville (since 2013). His main research areas are in labour economics and the economy of the electricity sector.

MARÍA MARTÍNEZ-BALLESTEROS received the M.Sc. degree in Computer Engineering and the PhD degree in Computer Science from the University of Seville, Spain, in 2012. Since 2009 she has been with the Department of Computer Science, University of Seville, where she is currently Associate Professor. Her primary areas of interest are data mining, machine learning techniques, association rules, evolutionary computation and Big Data.

JOSÉ C. RIQUELME received the M.Sc. degree in Mathematics and the PhD degree in Computer Science from the University of Seville, Spain. Since 1987 he has been with the Department of Computer Science, University of Seville, where he is currently Full Professor. His primary areas of interest are data mining, machine learning techniques, and evolutionary computation.

CARLOS USABIAGA BA (1988) and PhD (1992, with distinction) in Economics. Full Professor of Economics since 2003. Research stays, among other places, at Northwestern University (US) and the LSE (UK, in two occasions). Current Head of Doctoral Studies (since 2014) and former Head of the Department of Economics (2005-2013), both at Pablo de Olavide University (Seville). His main research areas are in macroeconomics and labour economics.

Capítulo 8

Big-Data Analysis for Demand Response in a Smart Electricity Market

Resumen

En este artículo se propone una nueva metodología para realizar un caracterización de los clientes de una compañía eléctrica a través de los datos de los consumos. Esta metodología aplicado a 1,8 TB de datos de una compañía eléctrica en una región europea. Los datos han sido generados por los contadores inteligentes instalados en los domicilios, y para realizar este análisis se han utilizado diferentes tecnologías de Big Data: HDFS para el almacenamiento de los datos distribuido; Apache Spark para el análisis se clustering con la versión distribuida de k-means; y para el cálculo del número óptimo de *clusters* se han usado los índices de validación internos para Big Data. El objetivo de esta metodología es la de conocer el comportamiento eléctrico de los consumidores para que las compañías eléctricas puedan adaptar sus tarifas a estos consumos, y viceversa. De esta manera, tanto eléctricas como consumidores lograrían eliminar los picos de consumo que existen en momentos puntuales.

- Estado: en revisión en IEEE Access (IEEE)
- Índice de Impacto (JCR 2018): 4.098
- Área de Conocimiento:
  - Engineering, electrical & electronics. Ranking 52/265 - Q1
  - Telecommunications. Ranking 19/88 - Q1
  - Computer Science, Information Systems. Ranking 24/155 - Q1

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.DOI

Big-Data Analysis for Demand Response in a Smart Electricity Market

J. A. FÁBREGAS[1*], J. M. LUNA-ROMERA[1*], D. GUTIÉRREZ-AVILÉS[2], J. C. RIQUELME[1]
[1]Department of Computer Science, University of Seville, Spain ({jfabregas,jmluna,riquelme}@us.es).
[2]Data Science & Big Data Lab, Pablo de Olavide University, ES-41013 Seville, Spain (davgutavi@upo.es).
[*]These authors contributed equally to this work

ABSTRACT The traditional business model of energy companies has undergone changes in recent years. The introduction of smart meters has led to an exponential increase in the volume of data available, whose analysis can aid in revealing consumption patterns among electricity customers to reduce costs and protect the environment. This information can help utilities use their facilities more efficiently, by avoiding major investment and expense in a capacity only used for a few hours a year. A set of techniques called *demand response* attempts to solve this issue using artificial intelligence proposals. This paper proposes a methodology for processing large volumes of data such as those generated by smart meters. For this analysis, big-data techniques are used, in particular a distributed version of the k-means algorithm, and four validation indices for the clustering for big data in Spark. In addition, a statistical analysis of the data is added. The original data corresponds to consumption by electricity customers in a European region for 7 years. The analysis of these consumers carried out in this work helps toward improving knowledge on consumption habits and types of customers.

INDEX TERMS Big Data, Smart Grids, Demand Response, Time-series Clustering, Electricity Consumption

I. INTRODUCTION

During most of the 20th century, the relationship between electricity users and distribution companies remained unchanged. Suppliers were not feely chosen, and therefore, there was no need to treat consumers as customers. However, deregulation, the green agenda, and continuous technological leaps have changed this relationship. New constraints, such as security of supply, competitiveness, and sustainability, now constitute the three priority axes towards changing the current energy model, which can be achieved though attaining objectives such as reducing emissions and improving renewable energy generation and energy efficiency.

An essential tool in this new model is provided by the so-called smart meters, which should not be understood only as devices that measure consumption but also as true sensors for an electrical network. These sensors facilitate a highly flexible and adaptable network that intelligently integrates the actions of the users connected to it, thereby achieving an efficient, safe, and sustainable supply.

One of the main problems in the electricity sector involves the need to provide generating capacity and an oversized network in order to cover peaks of high consumption of customers at specific times. However, there are now solutions based on adapting demand to available energy rather than increasing supply to satisfy demand. This is called Demand Response and aims to change customers' electricity consumption habits in response to changes in supply prices. In this way, companies can make better predictions and improve prices for customers without losing profits. Electricity networks therefore need to become an infrastructure that allows the flow of information between the participants in the electricity system. The main inconvenience is that the large volume of information available on these networks, can only be handled with big-data techniques.

Our proposal is based on the processing of this data in a parallel and distributed way, especially through the application of data-mining techniques to better understand the consumption patterns of electricity users. The distributed storage of data is performed in HDFS

[1] and the processing is carried out with Spark [2], a distributed and parallel computing platform. The k-means algorithm implemented in the Spark MLlib [3] library is used, as well as four clustering validation indices for big data [4].

In this paper, we focus on clustering, which is a data-mining method for grouping non-labelled instances from datasets. The idea is that the instances collected in the same group will have similar behaviour [5]. Particularly, in time-series clustering, this emerges as a useful approach towards mining frequent or uncommon interesting patterns from time-dependent data [6], [7], which is characterized by having high dimensionality and large size.

Furthermore, a statistical analysis of the tariffs contracted (power and access tariffs) by electricity consumers is carried out with the aim of ascertaining which tariffs are the best fit. The estimated average annual savings for these users are also calculated.

The results show the different types of consumers and the lack of adaptation between their consumption and their contracted tariffs. This paper could help in the planning of connections of renewable energy sources to the grid, with a twofold objective: price reduction, and environmental sustainability.

This paper is organized as follows. Section II describes the related work. Section III shows the characteristics of the dataset, while Section IV explains the preprocessing of the data. Section V details a statistical analysis of consumers, access tariffs, power, and prices. The experiments carried out with the application of clustering techniques are described in Section VI. And the conclusions of the study are presented in Section VII.

II. RELATED WORKS

The irregularity of electricity demand is one of the main issues in the sector, since power companies must have both oversized generation capacity and network redundancy in order to deal with large amounts of demand required for only a few hours a year. A threshold of 20% is usually established for the generation of latent electricity, which must cover approximately 5% of the network's service time (peak demand) [8]. A number of the resources to solve this problem are independent of the actions of customers, such as the development of new forms of storage, while others need the involvement of users for adequate demand management. These solutions are studied under the topic of *Demand Response* (DR) [9]. In contrast to conventional ideas of increasing supply to match demand, DR solutions aim to match demand with the energy available. In this case, the proponents of these types of measures take an active role in managing the demand by clients.

The objective of DR involves changing the patterns of energy consumption of customers in response to changes in the prices offered. This process allows electricity companies to better manage demand by better adjusting predictions and by reducing the cost of energy to customers. Multiple initiatives [10] of possible pricing schemes exists, whereby certain cases even maintain benefits for the supply companies [11]. One of the main advantages of DR is that it provides a sustainable option, especially in regions with a high presence of renewable generation sources, which are usually non-manageable (wind, fluid hydraulics, etc.).

In order to implement demand response mechanisms, electricity grids must evolve into an infrastructure that allows the flow of information between the different participants in the electricity system. In this field, big data becomes an essential technology for the analysis this flow of information to turn it into useful knowledge.

This customer consumption data, obtained from smart meters, is in the simple form of multiple time series. Time series can be understood as sequences of values observed over time in chronological order [12]. As time is a continuous variable, samples are recorded at equally spaced successive points. Therefore, each time series is a sequence of discrete time data.

In the context of time-series data mining, the major challenge is how to represent the time series data. The most common approach involves transforming the time into another domain for dimensionality reduction and developing an indexing mechanism. The similarity measures between time-series sub-sequences and segmentation constitute the two main tasks in time series mining that correspond to the classic tasks of data mining. The increasing use of time series-data has initiated a great deal of research and development in the field of data mining [13].

New concepts such as big data and its applications, have enabled research work on unsupervised solutions, such as clustering algorithms to extract knowledge from this avalanche of data. Clustering time series has been used to identify patterns that allow data analysts to extract valuable information from complex and massive datasets [5]. In [14], a combination of the Dynamic Time Warping (DTW) and Derivative Dynamic Time Warping (DDTW) distance measures were applied in an approach for hierarchical clustering of time-series data. In [15], a new algorithm for shape-based time-series clustering is proposed.

There are many applications of clustering techniques in a wide range of research fields, such as clustering gene-expression patterns in biology [16], analysing financial time series and the volatility of their returns in finance [17], detecting brain activity in medicine [18], and analysing temporal performance profiles of Unmanned Aerial Vehicles (UAV) [19].

There are many proposals in the field of clustering of energy consumption data, : In [20], the effect of similarity measures on the application of clustering to discover energy patterns in buildings is presented. To

obtain typical customer load profiles, a stability index is proposed in [21] to choose the grouping algorithm that best suits this pattern recognition problem. Moreover, another priority index (based on the stability index) is proposed for the determination of the priority range of the groupings. In [22], a partition clustering technique is developed to extract useful information from electricity prices. The same authors used clustering techniques in [23] to group and label samples in a dataset to predict the behaviour of time series based on the similarity of pattern sequences.

A new clustering framework for the automatic classification of loads of the electricity consumers is proposed in [24]. In [12], the authors propose an approach to assign the appropriate tariff to customers from the same specific group and different consumption behaviour based on some objective factors.

Regarding an intelligent management of electricity demand, in [25], is proposed a Virtual Power Player that manages and aggregates the available demand response and distributed generation resources in order to satisfy the required electrical energy demand and reserve. In [26], the k-means clustering algorithm and regression analysis where applied in order to determine the shape and the optimal number of seasonally-resolved residential demand profiles, as well as to draw correlations between the different profiles based on survey data from the occupants of the homes. An analysis of customer smart-meter data to better understand peak demand and the main sources of variability in customer behavior is presented in [27]. The aim of these authors was to identify suitable candidates for demand response and to improve the modelling of the energy profile.

In addition to classic data management methods, the big-data approach has recently emerged due to the availability of large amounts of data, distributed file systems, and powerful distributed processing engines. This has led to many of the data-mining algorithms, such as clustering algorithms, being adapted to the big-data environment. In the field of energy consumption, several major data solutions have now emerged, such as predictive analysis of real-time energy management in the field of energy consumption in [28], smart grid optimization in [29], and energy consumption patterns in [30].

III. DESCRIPTION OF THE DATASET

The original data was stored in 30 tables, divided into hundreds of CSV files. These tables contained a vast quantity of information on 4 million electrical customers from an important region of an European country over a period of 7 years (from 2010 to 2016)[1]. This information includes data on hourly consumption, tariffs, and meters with a total size of 1.8TB.

In the process of contracting an electricity tariff, consumers must choose an access tariff and a power rating. From the many access tariffs, we focus on only two: the uniform price rate (UPR), and the time-based pricing rate (TBPR) for consumers with less than 10 kW of contracted power. The vast majority of homes and small businesses are located within this power segment. UPR maintains a fixed price throughout the year while TBPR has a time discrimination of two periods. The peak and off-peak periods mark the two different prices depending on the date and time the energy is consumed. Peak period is from 12 h to 22 h in winter and from 13 h to 23 h in summer, while the off-peak period runs from 22 h to 12 h in winter and from 23 h to 13 h in summer. In this power segment is where the vast majority of homes and small businesses are located.

IV. DATA PRE-PROCESSING

This section explains the following stages of the data preprocessing: The data formatting and storage in distributed systems is described in Subsection IV-A. The feature selection process is explained in Subsection IV-B. The instance selection is detailed in Subsection IV-C. The time-series construction is defined in Subsection IV-D. Lastly, feature generation is explained in Subsection IV-E.

A. DATA FORMATTING AND STORAGE

The first objective is to store and process the raw data obtained. These raw files must be stored on our storage system. For this task, a Hadoop Distributed File System (HDFS) architecture has been installed and configured in our cluster. When HDFS takes the data, breaks the information down into separate blocks, and distributes these blocks to different nodes in the cluster. This provides great speed of access to data, as well as high fault tolerance thanks to the replicas this file system generates. For a better performance in a big-data context, CSV files are compressed into Parquet files. Parquet is a column-oriented data format with a highly efficient compression and coding scheme, which leads to low data storage cost and great efficiency for queries. The data is also stored in S3, the Amazon online storage system, to enable it to process large volumes of information with Amazon Web Service (AWS) tools.

B. FEATURE SELECTION

Once the data is stored in our systems, it is necessary to analyse the most important characteristics when carrying out a study of customers and their consumption patterns. The contracted access tariff and power are the main needed features required to build our dataset, along with the consumption measurements. This information is stored in Table 1, where the elements of the

[1]For commercial reasons, the authors can not reveal the origin of the datasets

processed tables are shown in units of million, and their size in megabytes.

TABLE 1: Processed tables

Table	Elements	Size
Customers	20.6	716
Contracts	40.6	560
Contract Master	33.1	666
Load Curves	2,094.44	340,992

These tables are processed to build a first dataset with the data of all customers who contracted less than 10 kW, and either UPR or TBPR access tariffs. This first dataset contains 853,977,202 instances with the following 27 features:

- Access tariff: the access tariff contracted by the customer. The only two possible values are UPR or TBPR.
- Power: the power contracted by the customer. This value is between 0.1 and 10 kW.
- Date: the date of the smart meter measurement.
- Hourly consumption: the measurement of electricity consumption for each hour of the day. A total of 24 measures.

Therefore, in this dataset there is one instance for each customer and day with the associated electricity consumption measurement.

C. INSTANCES SELECTION

The second step in the construction of a minable dataset is to select the instances that meet certain requirements. Firstly, users with a contracted power of 10 kW or less are selected. They are then filtered to include only those who have measurements for all 365 days of the year of 2016. Since instances are daily, we discard all instances that do not belong to those users. The number of instances after having applied these filters is 47,829,235. Finally, all the instances with null values in their consumption measures are also discarded. At this point, our dataset is composed of 47,829,235 daily instances. The instance structure is illustrated in Figure 1.

D. TIME-SERIES CONSTRUCTION

After selecting the optimal instances and features, a time-series dataset based on hourly customer consumption throughout 2016 is needed. For each customer, an instance is exists with its 24 measurements of hourly consumption for each day of 2016. Our objective is to transform these 365 instances for each client into a single instance. Therefore, the 24 measures of the 365 days are joined in a single row (see Figure 2).

1) Time series of consumption

Once the time series is built, the new instances are made up of 8,760 hourly measurements, in addition to the aforementioned features: power and access tariff. Finally, we discard those instances where all the consumption measurement values are 0. This dataset of users' electricity consumption contains 108,737 instances with 8,763 features.

2) Time series of normalized differences

The analysis of consumption patterns obtained directly from demand curves suffers from the limitation that customers tend to be grouped only in terms of their consumption. In order to study other patterns from data others than that of the total consumption, the data must be transformed. One of the most common transformations is involves obtaining the series of differences between the consumption of consecutive hours. Moreover, to further reduce the influence between high-consumption and low-consumption customers, the series of differences has been normalized by dividing its values by the average daily consumption of each customer. Therefore, the analysis of this transformed series provides information on patterns of increases or decreases in consumption over time in a non-dimensional way in relation to the size of demand.

The application of clustering techniques to the dataset of consumption and normalized differences together with an unsupervised analysis enables the consumer demands to be categorized.

E. FEATURE GENERATION

One of the main objectives in this paper is to carry out a statistical analysis of consumption and tariffs. To this end, the generation of new features is essential. First, the consumption and costs of energy in peak/off-peak hours need to be calculated according to the prices of the different markets. Subsequently, the annual consumption and costs of each consumer can also be calculated.

The other objective is to find consumption patterns among consumers. For this purpose, the contracted powers were categorized due to their wide variety. All these possible values were grouped into 8 segments. These segments are based on the standard power values that can currently be contracted trough the use of smart meters (Table 2):

As seen in Table 2, all powers below 2.3 are grouped in segment 1, since 2.3 is usually the lowest power contracted for households or commercial establishments. The powers between 0.345 and 1.725 are generally contracted to maintain basic electricity services in uninhabited places. For our study, 2.3 kW is considered as the minimum power to be contracted.

V. STATISTICAL ANALYSIS

In this section, a set of statistical analyses is performed. Subsection V-A presents an analysis of all the consumers. Subsection V-B includes a study of the contracted access tariffs. Subsection V-C shows an study

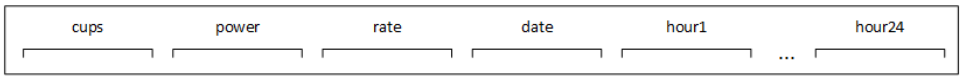


FIGURE 1: Daily instance

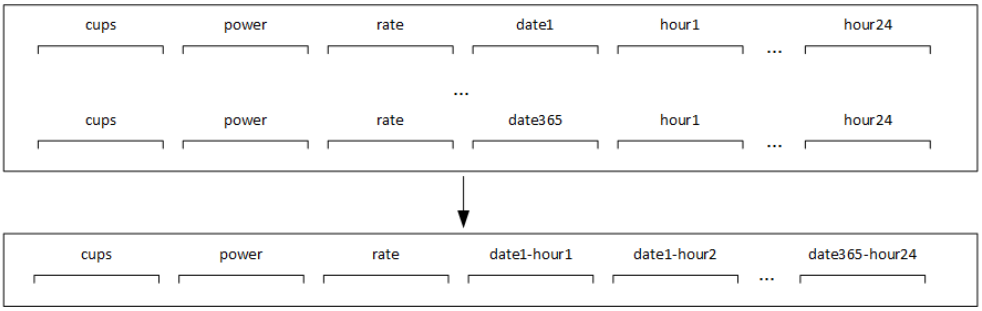


FIGURE 2: Daily to yearly instance transformation

TABLE 2: Power segments

Contracted Power(kW)	Power segment
(0-0.345)	1
[0.345-0.690)	1
[0.690-0.805)	1
[0.805-1.150)	1
[1.150-1.725)	1
[1.725-2.3)	1
[2.3-3.45)	2
[3.45-4.60)	3
[4.60-5.75)	4
[5.75-6.90)	5
[6.90-8.05)	6
[8.05-9.20)	7
[9.20-10.0)	8

TABLE 3: Customer distribution

(a) By contracted access tariff

Access tariff	Customers
UPR	102,123
TBPR	6,614

(b) By contracted power

Power(kW)	Customers
[0.345-2.3)	8,972
[2.3-3.45)	21,969
[3.45-4.60)	38,578
[4.60-5.75)	22,328
[5.75-6.90)	8,772
[6.90-8.05)	3,134
[8.05-9.20)	4,946
[9.20-10.0)	82

of the contracted powers. Lastly, an approximation of the annual savings of the consumers is calculated in Subsection V-D.

A. CUSTOMER ANALYSIS

All of the 108,737 electricity consumers included in this study contracted UPR and TBPR access tariffs and less than 10 kW. Table 3a shows the distribution of these clients according to their contracted access tariff. Table 3b displays the distribution of the clients for the various power segments defined in the previous subsection.

As seen in Table 3a, almost 94% of users chose the uniform access tariff (UPR) over the time-based pricing access tariff (TBPR). Moreover, as can be observed in Table 3b, approximately 75% contracted power between 2.3 and 5.75 kW. This indicates that the majority of these electricity customers contracted access tariffs with a fixed price and low to medium power.

B. CONTRACTED ACCESS-TARIFF ANALYSIS

There are two electricity markets: the free market and the price market. The prices of the free market are set by the electricity companies for the whole year, While in the regulated market there are maximum annual prices established by the government. The prices of this latter market change hourly and daily depending on the balance of supply-demand between whoever is producing energy (the generation company) and whoever is selling this energy to consumers (the company selling the electricity). The tariffs of this market are denominated voluntary small-consumer price (VSCP) tariffs. Users can therefore choose between either knowing what they pay for the whole year or adjusting to the daily prices depending on the market they select. Moreover, uniform price tariff and a time-based pricing tariff exists in both markets.

The approximate costs of the energy consumption of the different tariffs for each customer are calculated based on the prices of the free-market tariffs and the average prices for the VSCP tariffs. In order to analyse how the tariffs are adapted to the needs of consumers, these calculations are based on the prices of the year 2017. Table 4 shows the different prices of energy in each of the tariffs. Since the clients are final consumers, the value-added tax and the electricity tax have been added

to the price of the access tariffs.

TABLE 4: 2017 tariff prices

	Prices (€/kWh)		
	Uniform	Peak	Off-peak
Free market	0.14932	0.183754	0.08302
VSCP	0.15464	0.18163	0.08687

The tariff for the minimal annual cost to each customer is recommended. Table 5 shows the contracted tariffs int the rows and all possible tariffs to be contracted in the columns. In this way, the percentage of customers of a tariff who are advised to change to each of the other tariffs is shown. The row UPR corresponds to those users who contracted a free-market uniform tariff and the TBPR corresponds to those who contracted the time-based pricing tariff.

TABLE 5: Recommended tariff

	Free market		VSCP	
	UPR	TBPR	Uniform	Time
UPR	2.8%	96.2%	0%	1%
TBPR	1.1%	98.4%	0%	0.5%

As shown in Table 5, 96.2% of the users who contracted the UPR access tariff would find it more profitable to move to the TBPR. Only 2.8% of these customers should maintain their access tariff. Moreover, 98.4% of the users with TBPR have contracted the tariff that suits them the best. Therefore, TBPR remains the most recommended tariff for the vast majority of the electricity consumers analysed.

There are a 1% of clients with UPR and 0.5% of clients with TBPR who would save money changing to a VSCP time-based pricing tariff. Both groups of customers consumed more than 65% of electricity in the peak periods and less than 35% in the off-peak periods. In addition, the average contracted power of these customers is 5kW.

C. CONTRACTED POWER ANALYSIS

In terms of contracted power, Table 3b shows the distribution of the consumers grouped by power segments. As mentioned in the previous subsection, the most contracted powers were those between 2.3 and 5.75. The aim at this point was to ascertain whether the electrical consumers needed the power they had purchased or whether they could go down to lower power levels and therefore save money.

In order to establish the maximum peaks of power consumed, we have used the hourly consumption of 2016. In this way, as carried out in [31], [32], the demand at each instant t is calculated as the value measured at the time t plus a value in the form of white Gaussian noise, considered as being distributed according to a normal distribution $N(0, \sigma)$. Thus, it is possible to generate consumption density functions and to establish a limit value, in our case the 90th percentile (the probability that the maximum power reached is greater than that provided is less than 0.1), of the normal distribution considered here as the typical maximum consumption value in one hour. This maximum value is then compared to the maximum power of the segment previous to that contracted. In this way, it is possible to ascertain whether the client has contracted more power than necessary, and should therefore reduce said power.

For example, if a customer with a contracted power of 4.4 (segment [3.45-4.60)] had a peak consumption of 3.2 (segment [2.3-3.45)), its power could be lowered to 3.45.

The deviation σ has been calculated as the deviation of the consumption of each client throughout the entire sample. Therefore, we will estimate the maximum recorded consumption of a client for the 8,760 possible measurements and estimate its equivalent when taking the 90th percentile for the normal distributions that would generate said measurements. The probability that it is greater than the estimated values is less than 0.1.

After making this calculation, the results show that 59.6% of the consumers never exceeded their contracted power. These customers could therefore reduce their power without worrying about blowing the fuses due to an overly high consumption.

The power consumption peaks were also compared to the contracted power. In this case, the results obtained indicate that 17.47% of customers with less than 9.2 kW of power contracted have exceeded their contracted power at least once.

Figure 3 shows in terms of intervals on the x-axis the number of times customers exceeded their contracted power. The number of clients is shown on the y-axis. Finally, the line graph shows the cumulative percentage.

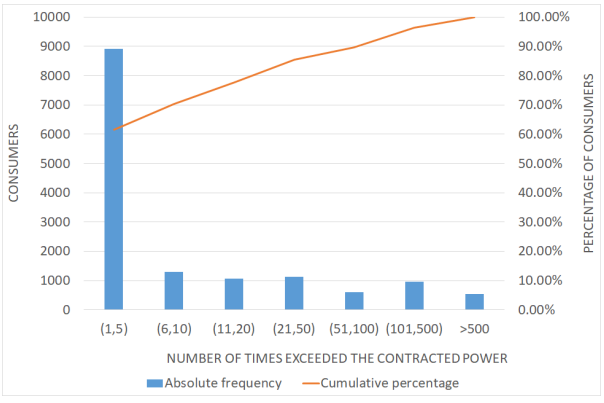


FIGURE 3: Customers who exceeded their contracted power

As seen in Figure 3, more than 70% of customers exceeded their contracted power up to 10 times, while

less than only 4% exceeded their contracted power on more than 100 occasions. It is important to note that the number of measurements is 8,760.

D. ESTIMATED SAVINGS

Once the annual costs have been calculated for each customer, the following step is to estimate their potential savings in terms of access tariff and power. In the previous subsection, the percentages of users who should change and those who should maintain their access tariffs were shown (5). In this subsection, Table 6 displays the estimated average yearly annual savings for these electricity customers. These savings are calculated by obtaining the difference in annual cost between the contracted access tariff and the recommended access tariff for each consumer. These calculations include value-added tax and electricity tax.

TABLE 6: Estimated average access tariff savings for 2017

	Free market		VSCP	
	UPR	TBPR	Uniform	Time
UPR	-	34 €	0 €	2 €
TBPR	8 €	-	0 €	0 €

As seen in Table 6, the biggest savings would be made for users who have to change from UPR to that of TBPR, which, as shown above, accounts for 96.2% of users of the uniform access tariff of the free market.

In access tariffs, the savings remain insignificant in economic aspects. The main savings are established in terms of power. The calculation of these savings are made as follows:

$$Savings = (kWp * pd * d) \quad (1)$$

where kWp is the daily price of the power, pd is the difference between the values of the different power segments, and d is the number of days. Table 7 displays the standard power values in rows and columns, and the rest of values are the estimated savings gained by reducing the contracted power between them. These calculations include the 21% of value-addex tax and the 5.11% of electricity tax.

As seen in Table 7, the savings by reducing the contracted power can be highly significant. As mentioned in the previous subsection, 62% of consumers could save money by contracting a lower power. Moreover, 96.2% also could save even more money by changing to the time-based-pricing access tariff. Therefore, the estimated savings for these clients would lie between 94 and 398 euros per year.

The estimated economic savings for each consumer have also been calculated in terms of hiring the most appropriate access tariff and power for their consumption habit (see Figure 4).

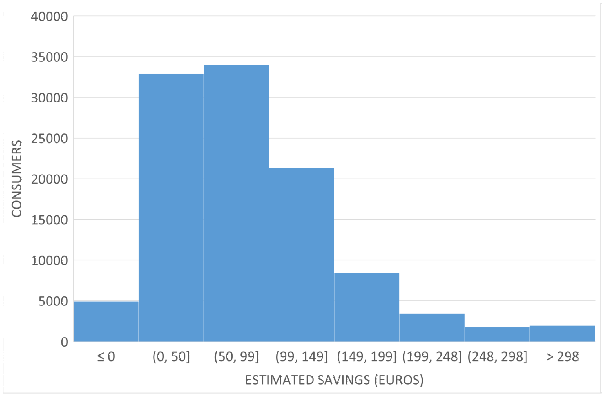


FIGURE 4: Estimated savings by number of consumers

Figure 4 shows how approximately 51,000 (over 47%) consumers could each save between 50 and 150 euros per year, and that about 15,000 (almost 14%) of them would save more than 150 euros.

All estimates of these savings are based exclusively on the price of energy (kWh) and price of power (kW) in each of the access tariffs and markets. No additional services or discounts have been considered. It is important to highlight that this analysis has been carried out without considering possible changes in the consumption habits of electrical customers. In that case, the savings would be even higher.

VI. EXPERIMENTS CONDUCTED

This section presents the results of several experiments conducted by using clustering techniques on the two datasets: consumptions and normalized differences. These experiments have been carried out in two different environments:

- Local cluster:
  - 72 processing cores: 64 Intel (R) Xeon (R) E7-4820 CPUs @ 2.00GHz, and 8 Intel (R) Core (TM) i7-7700K CPUs @ 4.20GHz.
  - 3 GeForce GTX 1080 GPUs with 2560 cores, Nvidia CUDA and 8 GB GDDR5X memory each.
  - 128 GB RAM: 64 GB DD3 and 64 GB DDR4.
  - 8 TB storage capacity.
  - Nodes interconnected through a Gigabit Ethernet network with a bandwidth of 1 Gbit/sec.
  - Hadoop HDFS 2.8.0.
  - Apache Spark framework 2.2.0.
- AWS EMR (Elastic Map Reduce) hardware:
  - Five instances of m3.2xlarge with Intel Xeon E5-2670 v2 (Ivy Bridge) processors with 16 CPUs, 30 GB RAM, and 2 SSDs of 80 GB each.

The rest of the section is organized as follows. Subsection VI-A presents the results of four clustering validity ratings to find the optimal k. Subsection VI-B details

TABLE 7: Estimated savings by reducing the contracted power

Power (kW)	2.3	3.45	4.60	5.75	6.90	8.05
3.45	60.58 €	-	-	-	-	-
4.60	121.16 €	60.58 €	-	-	-	-
5.75	181.74 €	121.16 €	60.58 €	-	-	-
6.90	242.32 €	181.74 €	121.16 €	60.58 €	-	-
8.05	302.90 €	242.32 €	181.74 €	121.16 €	60.58 €	-
9.20	363.48 €	302.90 €	242.32 €	181.74 €	121.16 €	60.58 €

the results of the clustering analysis obtained by the k-means. Subsection VI-C shows an evaluation of the clustering results. Finally, the results of the calculation of the estimated savings by cluster are presented in Subsection VI-D.

A. DETERMINING THE OPTIMAL NUMBER OF CLUSTERS

Before applying clustering algorithms to our datasets it is necessary to determine the optimal number of clusters (k) to obtain. To achieve this end, we applied four clustering validation indices for Big-Data (BD-CVIs) [4] to each of the datasets: BD-Silhouette [4], BD-Dunn [4], Davies-Bouldin [33], and Within Set Sum of Square Errors (WSSSE) [34].

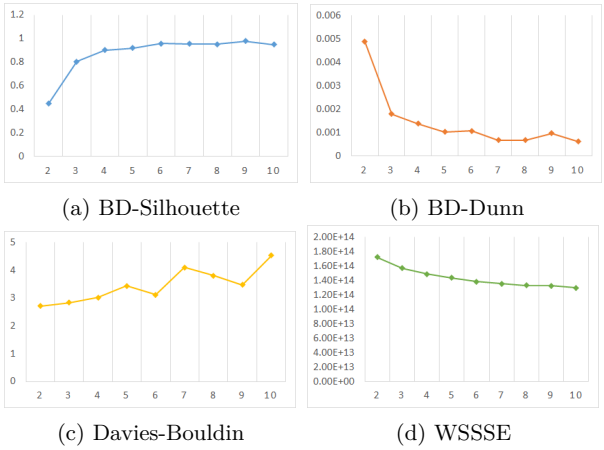
Figure 5a shows the graphical representation of the BD-Silhouette index results. For this index, the optimum values of k are their maximum, 6 and 9. These values match the maximum values in the graph corresponding to the BD-Dunn index (Figure 5b). The optimum values in the Davies-Bouldin index are found in the minimums. These coincide again at 6 and 9, as shown in Figure 5c. Finally, the results of the WSSSE index represented in Figure 5d fail to give a clear optimal value. In this index, a stabilization of values is sought and no specific value is found. After analyzing all these results, we have obtained the values 6 and 9 as optimal for the application of the k-means algorithm.

As for the consumption dataset, we have again applied these BD-CVIs to the dataset of normalized differences presented in SubsectionIV-D2. In this case, the results for the optimal values of k were 5 and 7.

B. CLUSTERING RESULTS

Once the optimal number of clusters has been calculated, the version implemented in Spark of the k-means algorithm is applied to each dataset. This implementation was developed to extract patterns in parallel and distributed systems. When running the algorithm, the Resilient Distributed Dataset (RDD) object and the previously obtained k are given as input. As a result, k clusters composed of the elements of the dataset are obtained.

In the electricity consumption dataset, two of the clusters obtained with k=9 had fewer than 5 elements, so we decided to work with the other optimal value obtained, k=6. In the dataset of normalized differences,



(a) BD-Silhouette (b) BD-Dunn

(c) Davies-Bouldin (d) WSSSE

FIGURE 5: BD-CVIs results for electricity consumption dataset

the value k=7 was selected since two of the clusters for k=5 contained a single element.

The distribution of the elements in the 6 clusters of the consumption dataset is shown in Table 8a, whereas Table 8b displays how the elements corresponding to the 7 clusters of the dataset of normalized differences are distributed.

TABLE 8: Clusters obtained for the datasets with optimal k

(a) Clusters obtained for the consumption dataset with k=6

Cluster	Elements
0	12,029
1	50,643
2	1,002
3	138
4	1,116
5	43,789

(b) Clusters obtained for the dataset of normalized differences with k=7

Cluster	Elements
0	42,276
1	8,241
2	2,544
3	18,994
4	28,462
5	7,191
6	1,029

1) Analysis of the clustering of the consumption dataset

In this sub-section, the results obtained when applying k-means to the dataset of electricity consumption are analyzed.

As seen in Table 8a, clusters 1 and 5 are made up of more than 86% of the consumers. Cluster 0 also contains a significant number of clients, while clusters 2, 3 and 4 are considerably smaller.

In order to show the characteristics of these clusters, their centroids have been graphically represented during a 7-day period in Figure 6a and another 24-hour period in Figure 6b.

Figure 6a shows the hourly consumption curves formed by the centroids of each of the clusters for a week in January 2016. It highlights that most of the consumers, grouped in clusters 1 and 5, have low consumption: under 1 kWh. As can be observed, the small group of customers which makes up cluster 3 has a very high consumption at night and virtually no consumption during the day. Cluster 2 is the only cluster where there is a clearly different pattern of consumption between workdays and weekends.

Figure 6b shows the consumption curves on a daily basis for a more detailed analysis. These measurements are from 11 January, 2016, coinciding with the first day of the week in the previous figure. As can be observed, the highest consumption peaks in clusters 0, 1 and 5 (more than 97% of customers) occur between 21 h and 22 h. Clusters 2 and 4 have two much more accentuated peaks of consumption. Those in cluster 2 at midday and late afternoon, and those in cluster 4 at 8 h and midnight. Finally, the elements of cluster 3 show a uniform consumption from 19 h to 8 h.

In order to clarify the differences in the consumption curves during the year, the centroids have been represented for one week of summer: from 11 to 17 July.

In Figure 7a, the main difference in the consumption curves is found in Cluster 4. The consumption of consumers in this cluster is considerably lower during the summer and its curve is almost identical to that of Cluster 0 at this time. The rest of the clusters have similar consumption behaviour throughout the year.

Figure 7b shows the differences between winter and summer electricity consumption in more detail. Clusters 0, 1, 3 and 4 have their highest consumption peaks between 22 h and 12 h. The curves of Cluster 3 are also noteworthy where consumption hours are reduced from 13 to 6 (12 h to 6 h).

2) Analysis of the clustering of the dataset of normalized differences

In this sub-section the results obtained when applying clustering to the dataset of normalized differences are analyzed.

The clusters of normalized differences have been graphically represented during the same periods as the consumption clusters seen above. In this case, the objective is to visualize the variations in consumption of the customers with respect to their daily average, without taking into account the amount of energy consumed.

Figure 8a shows how the highest peaks of consumption differences between hours belong to the elements of clusters 1, 2, 5 and 6. These clusters are the least numerous, and account for 17% of all consumers. This indicates that the consumption of the majority of users throughout the day maintains a certain uniformity.

At tweekends, the peaks of difference are gradually reduced in clusters 2 and 5. Meanwhile, cluster 1 shows one type of behaviour from Monday to Friday and another drastically different behaviour on Saturday and Sunday. However, the curves of cluster 6 remain almost the same throughout the week.

Figure 8b displays similar behaviour with a difference of one hour between the elements of clusters 1 and 5. This also occurs with clusters 3 and 4, which have peaks of difference in three periods of the day. Except for clusters 2 and 6, the rest have peaks early in the morning and between 19 h and 21 h.

The centroids have also been represented graphically in Figure 9a for the week of 11 to 17 July, 2016.

As can be observed in Figure 9a, the peaks of difference in clusters 1 and 5 are much smaller in summer than in winter. The rest of the clusters maintain similar behaviour during both seasons of the year.

Figure 9b shows the normalized differences over the course of 11 July. The main difference with respect to winter consumption is that the peaks are between 20 h and 23 h. In addition, the normalized differences between hourly consumption are smaller during the first hours of the day in all the clusters except for 2.

C. ANALYSIS OF RESULTS

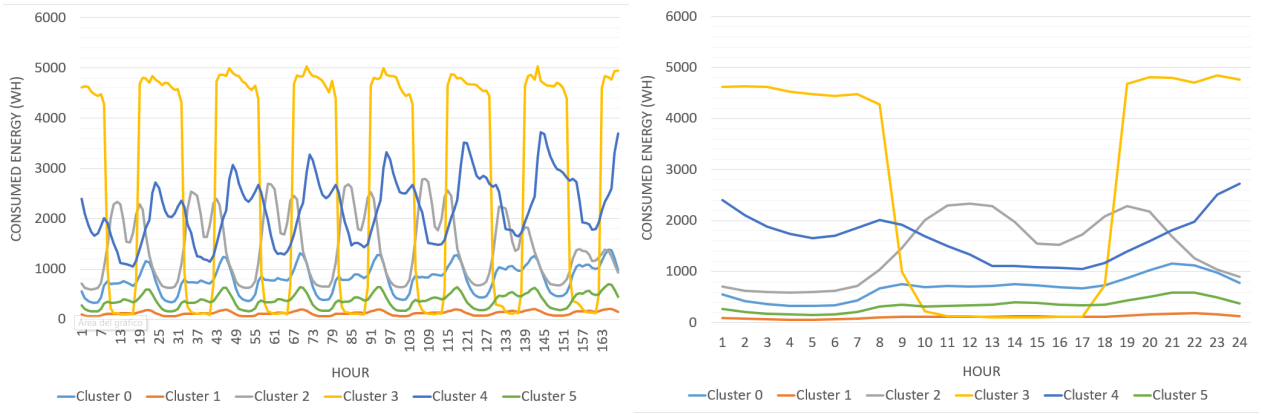
With the aim of categorizing the various groups of consumers, this section carries out an analysis by joining the results obtained in the previous section. For this purpose, contingency tables that show the correlation between the elements of the two clusters are generated.

In addition, although clustering is considered an unsupervised technique, a clustering validity analysis has been performed using the feature power segment. These contingency tables are generated by comparing the elements of the consumption clusters and those of the different power segments.

1) Unsupervised analysis

Table 9 shows the 6 clusters of the consumption dataset (C0 to C5) in the rows, and the 7 obtained from the normalized differences dataset in the columns,. These values represent the percentages relative to the total of each row, and hence the percentage of consumers of each consumption cluster present in each of the normalized difference clusters.

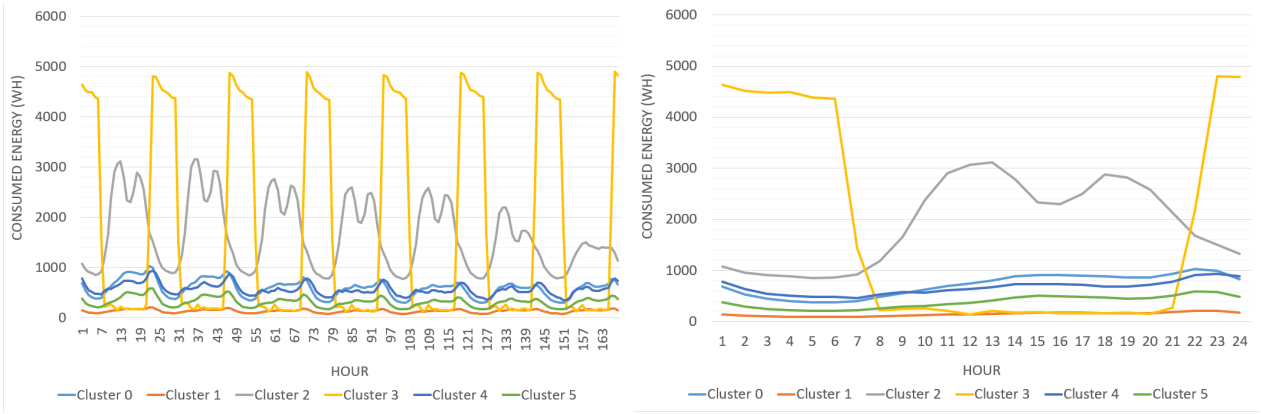
As shown in Table 9, almost the entire C3 cluster is made up of consumers in the D6 cluster. According to the Figure below, this would characterize a group of customers with very high night-time consumption. As



(a) Week of 11 to 17 January 2016

(b) January 11, 2016

FIGURE 6: Centroids of the consumption dataset in January 2016



(a) Week of 11 to 17 July, 2016

(b) July 11, 2016

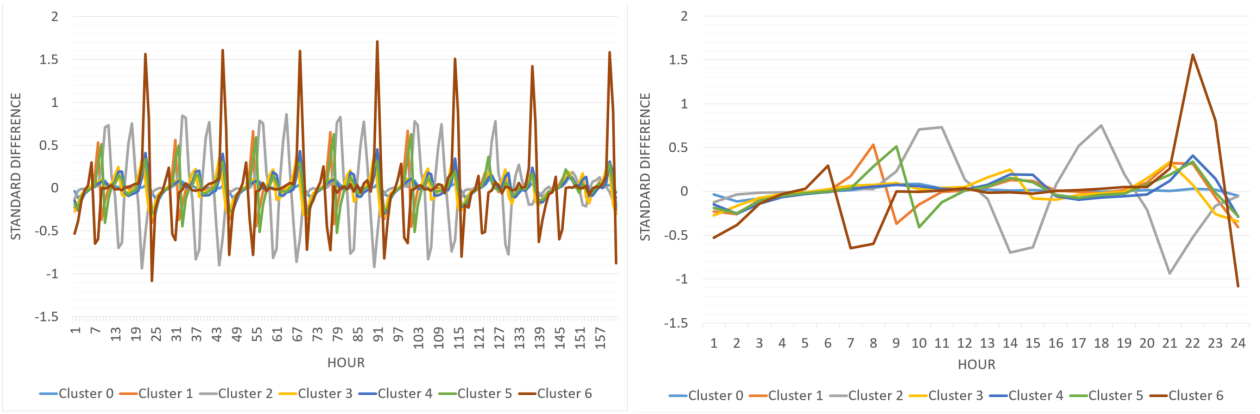
FIGURE 7: Centroids of the consumption dataset in July 2016

(a) Week of 11 to 17 January 2016

(b) January 11, 2016

FIGURE 8: Centroids of the standard difference dataset in January 2016

seen in Figures 6a and 7a, they have a minimum and a maximum in time differences that remain constant 7 days a week. This could indicate a very specific consumer profile, with high and uniform night-time energy



(a) Week of 11 to 17 July 2016

(b) July 11, 2016

FIGURE 9: Centroids of the standard difference dataset in July 2016

TABLE 9: Contingency table of values relative to the total of each row. Consumption clusters are showed in the rows and normalized differences clusters are displayed in the columns

	D0	D1	D2	D3	D4	D5	D6	
C0	29.11%	10.20%	3.85%	19.03%	27.53%	9.85%	0.43%	100.00%
C1	53.06%	5.25%	1.87%	13.03%	20.61%	4.74%	1.44%	100.00%
C2	52.20%	1.60%	27.64%	7.98%	7.29%	3.19%	0.10%	100.00%
C3	5.07%	0.00%	0.00%	0.00%	0.00%	0.72%	94.20%	100.00%
C4	56.16%	7.83%	0.09%	7.13%	16.99%	8.36%	3.43%	100.00%
C5	24.51%	9.70%	1.96%	22.71%	32.99%	7.94%	0.18%	100.00%

consumption.

In addition, half of cluster C1, the largest of the consumer clusters, is made up of consumers from cluster D0, the largest of the differences. Here, a large group of users with low consumption and few changes can be identified. These consumers have their greatest variations in consumption at 8 h and 18 h in winter and at 9 h and 22 h in summer.

Table 10 shows the relative percentages of the total for each column. In this case, these represents the percentage of consumers of each cluster of normalized differences present in each of the consumption clusters.

It can be observed in Table 10 that all the clusters of normalized differences are largely composed of clusters C1 and C5. This can be considered logical, as these two clusters represent 86% of consumers. Clusters D0 and D1 contain the majority of C1 elements, while clusters D1, D3, D4 and D5 have a higher presence of C5 elements. In D2, clusters C1 and C5 do not have such a high representation due to the 10% of elements of the C2 cluster.

Focusing on cluster D6, 70.65% of its elements belong to cluster C1. While Table 9 shows that only 1.44% of those in cluster C1 of consumption belong to cluster D6. This indicates that a large proportion of consumers with high peaks of difference in the afternoons have low energy consumption, although there are very few customers with low energy consumption who have these high differences.

2) Semi-supervised analysis

The objective of crossing the results of the consumption clusters with the segments of contracted power is to find relationships between the different types of consumers and their contracted powers.

Table 11 shows the consumption dataset clusters per row (C0 to C5) and the power segments per column (P1 to P8). These values represent the percentages relative to the total of each row, that is, the percentage of consumers of each consumption cluster present in each of the power segments.

As shown in Table 11, for the clusters C1 and C5 (those of lower consumption), most of their elements belong to the ranges P1 to P4, the lowest powers. Meanwhile, in clusters C2 and C4, consumers have higher power ratings (P5, P6 and P7). As expected, most of the customers with low consumption have only contracted medium and low power. In addition, more than half of the customers with high consumption contracted high power. A particular case is C3, where almost half of its elements belong to P6. This indicates that approximately 50% of customers with high night-time consumption have contracted between 6.90 and 8.05kW.

In Table 12, the percentages relative to the total of each column are represented, that is, the percentage of consumers of each power range present in each of the

TABLE 10: Contingency table of values relative to the total of each column. Consumption clusters are showed in the rows and normalized differences clusters are displayed in the columns

	D0	D1	D2	D3	D4	D5	D6
C0	8.28%	14.89%	18.20%	12.05%	11.63%	16.48%	5.05%
C1	63.56%	32.29%	37.15%	34.74%	36.68%	33.39%	70.65%
C2	1.24%	0.19%	10.89%	0.42%	0.26%	0.45%	0.10%
C3	0.02%	0.00%	0.00%	0.00%	0.00%	0.01%	12.63%
C4	1.51%	1.08%	0.04%	0.43%	0.68%	1.32%	3.79%
C5	25.39%	51.55%	33.73%	52.36%	50.76%	48.35%	7.77%
	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

TABLE 11: Contingency table of values relative to the total of each row. Consumption clusters are showed in the rows and power segments are displayed in the columns

	P1	P2	P3	P4	P5	P6	P7	P8	
C0	2.65%	9.51%	27.46%	23.09%	17.67%	6.66%	12.80%	0.17%	100.00%
C1	13.76%	26.37%	34.08%	17.74%	4.43%	1.50%	2.08%	0.03%	100.00%
C2	2.69%	5.29%	12.28%	16.87%	14.87%	19.56%	26.85%	1.60%	100.00%
C3	3.62%	0.72%	10.87%	21.74%	10.14%	47.10%	4.35%	1.45%	100.00%
C4	4.15%	9.01%	12.54%	15.19%	26.50%	19.08%	12.90%	0.62%	100.00%
C5	3.67%	16.70%	40.50%	23.28%	8.89%	2.50%	4.41%	0.05%	100.00%

consumption clusters.

Table 12 shows how more than 85% of consumers in ranges P1 to P4 belong to clusters C1 and C5. Therefore, the relationship between these groups of clusters and power ranges exists in both directions: Customers with low consumption contracted a low or medium power level, and vice versa.

In the analysis of Tabla 11, it was shown how the high-consumption customers had high contracted power. However, regarding the ranges P5 to P8, they are also made up of between 45% and 70% of users with low consumption. Therefore, although customers with high consumption have high contracted power, the majority of consumers with these powers have low consumption.

Focusing on the case of P6 and C3. While 47.1% of C3 consumers were on P6, only 2.07% of P6 consumers are on C3. This reinforces the above conclusion, since more than 85% of consumers with power between 6.9 and 8.05kW consumes between 0.5 and 1.5kWh.

D. ESTIMATED SAVINGS

Section V-D presented the results of the calculations on the estimated savings for individual customers. These results were shown in Figure 4. After grouping and characterizing consumers according to their consumption as shown in Table 8a, new calculations of estimated savings have been carried out.

Table 13 shows the number of consumers belonging to each of the clusters that would save money by contracting the power and access rate that best suits their consumption. The average of these savings is also displayed.

The results presented in Table 13 show that more than 97% of consumers belonging to the largest clusters (C1 and C5) can save approximately 90 euros per year. The percentage of consumers in clusters C3 and C4 who could save is much lower, although, their average savings exceed 119 and 188 euros per year, respectively. As previously shown in Figures 6a and 7a, these clusters were those that showed the greatest differences in consumption patterns from the remaining clusters, as well as those that differed the most between summer and winter.

VII. CONCLUSIONS

This paper presents an analysis of the electricity consumers in an European region. We applied big-data techniques to consumption data with the aim of finding patterns in the behaviour of the consumers. The use of internal validation indices proved useful in obtaining an optimal number of clusters in the two datasets analysed. The statistical analysis and the results of the clustering evaluation showed that the time-based price access tariff of the free market (TBPR) is the best option for most users, after analysing their annual consumption. This is because due to the fact that through the year, consumption peaks in the mornings and at midday occur during off-peak hours. This is also true for night peaks in summer. However, only 5% of these clients have signed up for the time-based access tariff. The results also indicate that at least 59.6% of customers would save money on their bills by reducing the contracted power and just 17.47% of consumers exceeded their contracted power. The estimated savings for the consumers would exceed, on average, 90 euros per year by contracting the power and access tariff that best suits their consumption habits.

The results obtained can be useful for both electricity companies and their consumers. Companies can offer tariffs that are better adapted to the consumption of their customers, who can change their consumption patterns to obtain greater savings on their bills. In addition, from the point of view of the utilities, it can help towards improving investment planning and towards designing

TABLE 12: Contingency table of values relative to the total of each column. consumption clusters are showed in the rows and power segments are displayed in the columns

	P1	P2	P3	P4	P5	P6	P7	P8
C0	3.54%	5.20%	8.55%	12.42%	24.34%	25.53%	31.10%	24.39%
C1	77.66%	60.80%	44.75%	40.25%	25.72%	24.31%	21.33%	15.85%
C2	0.30%	0.24%	0.32%	0.76%	1.71%	6.25%	5.44%	19.51%
C3	0.06%	0.00%	0.04%	0.13%	0.16%	2.07%	0.12%	2.44%
C4	0.52%	0.46%	0.37%	0.77%	3.44%	6.89%	2.95%	8.54%
C5	17.92%	33.28%	45.98%	45.66%	44.63%	34.94%	39.06%	29.27%
	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

TABLE 13: Estimated savings by consumption clusters

Cluster	Customers	Potential savers	Potential savers(%)	Average savings
C0	12,029	4,081	33.93%	100.52 €
C1	50,643	49,703	98.14%	86.77 €
C2	1,002	891	88.92%	165.60 €
C3	138	19	13.77%	119.61 €
C4	1,116	511	45.79%	188.77 €
C5	43,789	42,213	96.40%	96.18 €

marketing strategies focused mainly on the customer clusters that present greater potential savings.

REFERENCES

[1] K. Shvachko, H. Kuang, S. Radia, and R. Chansler. The hadoop distributed file system. In 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pages 1–10, May 2010.

[2] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. 10:10–10, 07 2010.

[3] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar. Mllib: Machine learning in apache spark. Journal of Machine Learning Research, 17(34):1–7, 2016.

[4] José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros, and José C. Riquelme Santos. An approach to validity indices for clustering techniques in big data. Progress in Artificial Intelligence, 7(2):81–94, Jun 2018.

[5] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering – a decade review. Information Systems, 53:16 – 38, 2015.

[6] T. Warren Liao. Clustering of time series data — a survey. Pattern Recognition, 38(11):1857 – 1874, 2005.

[7] Haixun Wang, Wei Wang, Jiong Yang, and Philip Yu. Clustering by pattern similarity in large data sets. 3, 10 2002.

[8] Haider Tarish Haider, Ong Hang See, and Wilfried Elmenreich. A review of residential demand response of smart grid. Renewable and Sustainable Energy Reviews, 59:166 – 178, 2016.

[9] Ijaz Hussain, Sajjad Mohsin, Abdul Basit, Zahoor Ali Khan, Umar Qasim, and Nadeem Javaid. A review on demand response: Pricing, optimization, and appliance scheduling. Procedia Computer Science, 52:843 – 850, 2015.

[10] Qi Wang, Chunyu Zhang, Yi Ding, George Xydis, Jianhui Wang, and Jacob Østergaard. Review of real-time electricity markets for integrating distributed energy resources and demand response. Applied Energy, 138:695 – 706, 2015.

[11] Haider Tarish Haider, Ong Hang See, and Wilfried Elmenreich. Residential demand response scheme based on adaptive consumption level pricing. Energy, 113:301 – 308, 2016.

[12] Francisco Martínez-Álvarez, Alicia Troncoso, Gualberto Asencio-Cortés, and José C. Riquelme. A survey on data mining techniques applied to electricity-related time series forecasting. Energies, 8(11):13162–13193, 2015.

[13] Tak chung Fu. A review on time series data mining. Engineering Applications of Artificial Intelligence, 24(1):164 – 181, 2011.

[14] Maciej Łuczak. Hierarchical clustering of time series data with parametric derivative dynamic time warping. Expert Systems with Applications, 62:116 – 130, 2016.

[15] Xiaohang Zhang, Jiaqi Liu, Yu Du, and Tingjie Lv. A novel clustering method on time series data. Expert Systems with Applications, 38(9):11891 – 11900, 2011.

[16] Mikhail Pyatnitskiy, Ilya Mazo, Maria Shkrob, Elena Schwartz, and Ekaterina Kotelnikova. Clustering gene expression regulators: New approach to disease subtyping. PLOS ONE, 9(1):1–10, 01 2014.

[17] A. Stetco, X. j. Zeng, and J. Keane. Fuzzy cluster analysis of financial time series and their volatility assessment. In 2013 IEEE International Conference on Systems, Man, and Cybernetics, pages 91–96, Oct 2013.

[18] Martijn van den Heuvel, René Cw Mandl, and Hilleke E. Hulshoff Pol. Normalized cut group clustering of resting-state fmri data. PLoS ONE, 3:537 – 541, 2008.

[19] Víctor Rodríguez-Fernández, Héctor D. Menéndez, and David Camacho. Analysing temporal performance profiles of uav operators using time series clustering. Expert Systems with Applications, 70:103 – 118, 2017.

[20] Félix Iglesias and Wolfgang Kastner. Analysis of similarity measures in times series clustering for the discovery of building energy patterns. Energies, 6(2):579–597, 2013.

[21] T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang. A new index and classification approach for load pattern analysis of large electricity customers. IEEE Transactions on Power Systems, 27(1):153–160, Feb 2012.

[22] F. Martínez-Álvarez, A. Troncoso, J. C. Riquelme, and J. M. Riquelme. Partitioning-clustering techniques applied to the electricity price time series. In Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL'07, pages 990–999, Berlin, Heidelberg, 2007. Springer-Verlag.

[23] F. Martinez Alvarez, A. Troncoso, J. C. Riquelme, and J. S. Aguilar Ruiz. Energy time series forecasting based on pattern sequence similarity. IEEE Transactions on Knowledge and Data Engineering, 23(8):1230–1243, Aug 2011.

[24] Félix Biscarri, Iñigo Monedero, Antonio García, Juan Ignacio Guerrero, and Carlos León. Electricity clustering framework for automatic classification of customer loads. Expert Systems with Applications, 86:54 – 63, 2017.

[25] Pedro Faria, Zita Vale, and José Baptista. Demand response programs design and use considering intensive penetration of distributed generation. Energies, 8(6):6230–6246, 2015.

[26] Joshua D. Rhodes, Wesley J. Cole, Charles R. Upshaw, Thomas F. Edgar, and Michael E. Webber. Clustering analysis of residential electricity demand profiles. Applied Energy, 135:461 – 471, 2014.

[27] S. Haben, C. Singleton, and P. Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Transactions on Smart Grid, 7(1):136–144, Jan 2016.
[28] N. Balac, T. Sipes, N. Wolter, K. Nunes, B. Sinkovits, and H. Karimabadi. Large scale predictive analytics for real-time energy management. In 2013 IEEE International Conference on Big Data, pages 657–664, Oct 2013.
[29] Jui-Sheng Chou and Ngoc-Tri Ngo. Smart grid data analytics framework for increasing energy savings in residential buildings. Automation in Construction, 72:247 – 257, 2016.
[30] R. Perez-Chacon, R. L. Talavera-Llames, F. Martinez-Alvarez, and A. Troncoso. Finding electric energy consumption patterns in big time series data. In Sigeru Omatu, Ali Semalat, Grzegorz Bocewicz, Paweł Sitek, Izabela E. Nielsen, Julián A. García García, and Javier Bajo, editors, Distributed Computing and Artificial Intelligence, 13th International Conference, pages 231–238, Cham, 2016. Springer International Publishing.
[31] Neha Nandakumar. Computational models of natural gas markets for gas-fired generators. Massachusetts Institute of Technology, 2016.
[32] Arcos-Vargas. Ángel, José María Luna-Romera, Jorge García-Gutiérrez, and José C. Riquelme Santos. Smart meters: potential savings for consumers. DYNA, 2018.
[33] David L. Davies and Don Bouldin. A cluster separation measure. PAMI-1:224 – 227, 05 1979.
[34] Spark clustering rdd based api documentation for spark 2.3.0. 2017. urlhttps://spark.apache.org/docs/2.3.0/mllib-clustering.html, 2018. Accessed: 2018-06-30.

JOSÉ A. FÁBREGAS José A. Fábregas received the M.Sc. degree in Computer Engineering from the University of Seville, Spain. Since 2017 he has been with the Department of Computer Science, University of Seville, where he is currently a researcher. His primary areas of interest are data mining, machine learning techniques and Big Data.

JOSÉ MARÍA LUNA-ROMERA José María Luna-Romera is a full time PhD research student at University of Sevilla (Spain) since November 2015 after being awarded with a four-year research scholarship by Spanish Government. He received his M.Sc. degree in Software Engineering and Technology in 2012 and published his master dissertation on Data Mining Applied to Earthquakes Prediction. His current research interests concern clustering analysis and more generally data mining and Big Data.

DAVID GUTIÉRREZ-AVILÉS David Gutiérrez-Avilés received the M.Sc. degree in Computer Engineering and the Ph.D. degree in Computer Science from the University of Seville, Spain. Since 2015 he has been in the Data Science & Big Data Research Lab of the Pablo de Olavide University of Seville, where he is currently a researcher and assistant professor. His main research lines are focused on: Electricity fraud detection in Big Data environments, On-line machine learning from Big data streaming, Analysis of Internet of Things protocols, and sensor data analysis.

JOSÉ C. RIQUELME José C. Riquelme received the M.Sc. degree in Mathematics and the Ph.D. degree in Computer Science from the University of Seville, Spain. Since 1987 he has been with the Department of Computer Science, University of Seville, where he is currently Full Professor. His primary areas of interest are data mining, machine learning techniques, and evolutionary computation.

Parte III

Conclusiones y Trabajos Futuros

Capítulo 9

Conclusiones y Trabajos Futuros

Ahora hace falta recoger los trozos de prudencia,
aunque siempre nos falte alguno;
recoger la vida vacía
y caminar esperando que lentamente se llene,
si es posible otra vez, como antes,
de sueños desconocidos y deseos invisibles

Luis Cernuda
Telarañas cuelgan de la razón

9.1. Conclusiones

Esta tesis doctoral por compendio de artículos se ha desarrollado desde dos puntos de vista diferentes. El primero de ellos se basa en el desarrollo, diseño e implementación de nuevos índices de validación de clustering, dos índices internos especialmente diseñados para trabajar con Big Data, y un índice externo basado en el test estadístico chi cuadrado. Desde el segundo punto de vista podemos ver que estos índices se han aplicado a problemas reales, consiguiendo transferir los resultados a otros grupos de investigación o las empresas. A continuación, se detallan las conclusiones obtenidas de cada uno de estos puntos de vista:

- Se han diseñado dos nuevos índices de validación internos de clustering con el objetivo de poder trabajar con datos que podrían ser considerados Big Data. Estos nuevos índices, BD-Silhouette y BD-Dunn, han sido basados en índices tradicionales (Silhouette y Dunn). La característica principal respecto a sus versiones tradiciones, es la reducción de la complejidad algorítmica de los mismos, ya que se han simplificado sus implementaciones originales de cara a poder trabajar con grandes cantidades de datos.

- Por otra parte, se ha desarrollado un novedoso índice de validación externo de clustering basado en el test estadístico de chi cuadrado llamado Chi Index. Este nuevo índice ofrece un resultado de clustering directo sin necesidad de que su resultado sea interpretado. Su efectividad ha sido testada frente a otros 15 índices de la literatura quedando significativamente por encima de sus competidores en 47 datasets reales.

La segunda parte de la tesis ha sido la de aplicar estos nuevos índices en proyectos reales, haciendo uso de los datos proporcionado por diferentes fuentes. Se han aplicado estos índices en tres proyectos de investigación:

- Se ha implementado una metodología para realizar un análisis de clustering en datos de consumo eléctrico de la Universidad Pablo de Olavide. Esta metodología está preparada para ser aplicada en datos de consumos eléctricos de una *smart city* ya que se ha hecho uso de tecnologías y aplicaciones especialmente diseñadas para Big Data. En este proyecto se hizo uso de los índices de validación internos para Big Data, obteniendo unos resultados que podrían ser usados por las diferentes administraciones con la idea de optimizar el uso de la energía en los edificios de una *smart city*.

- En segundo proyecto real se han usado los datos de las colocaciones registradas por Ministerio de Trabajo, Migraciones y Seguridad Social. En este proyecto se ha realizado un análisis de *clustering* de dos periodos económicos diferentes, el 2011-2013, años de plena crisis económica; y 2014-2016, periodo de recuperación. Dado el tamaño de los datos de ambos periodos, 1,9 y 2,4 millones de colocaciones, se podría considerar un problema de Big Data. En este análisis se ha hecho uso de los índices internos de Big Data, así como del Chi Index para obtener la solución de clustering óptima en base a unas etiquetas establecidas como son la comunidad autónoma, la provincia, la actividad y la ocupación. Una vez obtenida la solución de *clustering* óptima se ha realizado una caracterización de los *clusters*, pudiendo hacer una comparativa entre los *clusters* resultantes de ambos periodos. Los resultados de este proyecto podrían ser de utilidad a la administración pública de cara a la toma de decisión en políticas de empleo.

- En el tercer proyecto se ha diseñado una nueva metodología para caracterizar a los consumidores de una compañía eléctrica en función de sus hábitos de consumo. La caracterización se ha realizado aplicando técnicas de *clustering* usando tecnologías de Big Data a 1,8 TB de datos. Las técnicas de *clustering*, así como su análisis se llevaron a cabo usando tecnologías como: HDFS, para el almacenamiento de los datos; y *Apache Spark* para los métodos de *clustering* y sus índices de validación. Los resultados de este estudio dan a conocer el comportamiento

de los consumidores con el fin de que las compañías eléctricas puedan adaptar sus tarifas a los consumos, así como los consumidores ajusten sus consumos a las nuevas tarifas con el fin de reducir los picos de consumo que existen. De esta forma se consigue optimizar la energía generada por las compañías.

9.2. Trabajo futuro

Como principal trabajo futuro se propone diseñar un nuevo algoritmo de *clustering* que basado en alguna heurística de optimización tome como función de fitness las medidas de calidad propuestas. Una segunda línea de trabajo es explotar estos índices de validación mediante técnicas de *clustering* jerárquico. Esta técnica de *clustering* proporciona como salida un dendrograma que puede variar según como se determine la distancia entre *clusters* (completa, mínima, máxima, etc). En estos dendrogramas se encuentran "todas"las posibilidades de un *clustering* óptimo. La idea es implementar una búsqueda del *cluster* óptimo sobre los dendrogramas usando de nuevo como fitness los CVI propuestos. Otra línea de investigación poco explotada en el aprendizaje no supervisado es la selección óptima de atributos. Al contrario de la versión supervisada, la selección de atributos en el ámbito del *clustering* ha tenido mucho menos interés por parte de la comunidad de Data Science. Sin embargo, no deja de ser importante seleccionar los atributos que mejor determinan el *clustering* óptimo. De nuevo, se propone la implementación de heurísticas de búsqueda que usando los CVI definidos como medida de fitness sean capaces de inferir subconjuntos de atributos que optimicen el *clustering*.

9.3. Conclusions

This doctoral thesis, presented as a compendium of articles, has been developed from two different points of view. The first is based on the development, design, and implementation of new clustering validation indices: two internal indices specially designed to work with Big Data, and an external index based on the chi-square statistical test. From the second point of view, we can see that these indices have been applied to real problems, and have enabled the results to be transferred to other research groups or companies. Below, the conclusions obtained from each of these points of view are detailed:

- Two new internal clustering validation indices have been designed with the aim of working with data that could be considered Big Data. These new indices, BD-Silhouette and BD-Dunn, have been built based on traditional indices (Silhouette and Dunn). The main improvement with respect to their traditional versions, lies in the reduction of their algorithmic complexity, since their original implementations have been simplified in order to be able to work with large amounts of data.

- On the other hand, an innovative external clustering validation index of has been developed based on the chi-squared statistical test called Chi Index. This new index offers a direct clustering result without the need for its result to be interpreted. Its effectiveness has been tested against 15 other indices in the literature, and achieves superior performance compared to all its competitors in 47 real datasets.

The second part of the thesis deals with the application of these new indices in real projects by making use of the data provided by different sources. These indices have been applied in three research projects:

- A methodology has been implemented to perform a clustering analysis on electricity consumption data of the Pablo de Olavide University. This methodology is prepared to be applied in data regarding the electricity consumption of a smart city since it has made use of technologies and applications specially designed for Big Data. In this project, the internal validation indices for Big Data were used, and results were obtained that could be used by the different administrations in order to optimise the use of energy in the buildings of a smart city.

- • In the second real project, the data on the placements registered by the Ministry of Labour, Migrations, and Social Security has been used. In this project, a clustering analysis of two different economic periods has been carried out: 2011-2013, which are years of full economic crisis; and 2014-2016, a recovery period. Given the size of the data in the two periods, 1.9 and 2.4 million placements, respectively, it could

be considered a Big Data problem. In this analysis, the internal Big Data indices have been employed, as well as the Chi Index to obtain the optimal clustering solution based on established labels such as the autonomous community, the province, activity, and occupation. Once the optimal clustering solution was obtained, a characterisation of the clusters was carried out, by making a comparison between the clusters that resulted from the two periods. The results of this project could be useful for the public administration regarding decision-making in employment policies.

- In the third project, a new methodology has been designed to characterise the consumers of an electric company based on their consumption habits. The characterization has been carried out by applying clustering techniques using Big Data technologies to 1.8 TB of data. The clustering techniques, as well as their analysis, were carried out through the use of technologies such as: HDFS, for the storage of the data; and Apache Spark for clustering methods and their validation indices. The results of this study reveal the behaviour of consumers and hence not only can electricity companies adapt their tariffs to consumption, but consumers can also adjust their consumption to fit the new tariffs, which lead to a reduction in consumption peaks. In this respect, it is possible to optimise the energy generated by the companies.

9.4. Future Work

As the subsequent main line of research, the aim is to design a new clustering algorithm based on an optimisation heuristic, whereby the proposed quality measures are taken as a fitness function. A second line of future research is to exploit these validation indices by using hierarchical clustering techniques. This clustering technique provides a dendrogram that can vary depending on how the distance between clusters is determined (e.g., complete distance, minimum distance, and maximum distance). These dendrograms include .[a]ll"the possibilities of an optimal clustering. The idea is to implement a search of the optimal cluster on the dendrograms while maintaining the proposed CVI as a fitness measure.

Another area of research which has not been fully exploited in unsupervised learning is that of the optimal selection of attributes. Unlike the supervised version, the selection of attributes in the scope of clustering has attracted little interest from the Data Science community. However, it remains crucial that the attributes that best determine optimal clustering are selected. Again, the implementation of search heuristics, with the use of the defined CVI as a fitness measure, could provide the subset of features that optimise clustering.

Parte IV

Apéndices

Apéndice A

Curriculum

A.1. Revistas indexadas JCR

1. Título: **A study of the suitability of autoencoders for preprocessing data in breast cancer experimentation**. Autores: **Laura Macías-García, José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros, José C. Riquelme, Ricardo González-Cámpora**.

   Publicado en: **Journal of Biomedical Informatics**, Elsevier, ISSN: 1532-0464, Fecha de Publicación: Agosto 2017, Volumen: 77, En Páginas: 33-44, DOI: `https://doi.org/10.1016/j.jbi.2017.06.020`, **Q2 en Computer Science, Interdisciplinary Applications (28/105). Q1 en Medical Informatics (7/25). JCR-2017 F.I.: 2.882**. Citas Scopus-Scholar-WOS: 4-4-3.

2. Título: **Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities**. Autores: **Rubén Pérez-Chacón, José María Luna-Romera, Alicia Troncoso, Francisco Martínez-Álvarez, José C. Riquelme**.

   Publicado en: **Energies**, MDPI, ISSN: 1996-1073, Fecha de Publicación: Marzo 2018, Volumen: 11, Número: 3, En Páginas: 683, DOI: `https://doi.org/10.3390/en11030683`, **Q3 en Energy and Fuels, (48/97). JCR-2018 F.I.: 2.707**. Citas Scopus-Scholar-WOS: 17-22-13.

3. Título: **¿Cómo transformar información en ahorro para el consumidor doméstico? El caso del contador eléctrico inteligente en España**. Autores: **Ángel Arcos-Vargas, Jose María Luna-Romera, Jorge García-Gutiérrez, José C Riquelme-Santos**.

   Publicado en: **DYNA**, ISSN: 0012-7361, Fecha de Publicación: Mayo 2018, Volumen: 93, en Páginas: 244, DOI: `https://doi.org/10.6036/`

8782, **Q4 en Engineering, Multidisciplinary (76/86). JCR-2017 F.I.: 0.520**.

4. Título: **External Clustering Validity Index based on chi-squared statistical test**. Autores: **José María Luna-Romera, María Martínez-Ballesteros, Jorge García-Gutiérrez, José C. Riquelme**.

   Publicado en: **Information Sciences**, Elsevier, ISSN: 0020-0255, Fecha de Publicación: Junio 2019, Volumen: 487, en Páginas: 1-17, DOI: https://doi.org/10.1016/j.ins.2019.02.046, **Q1 en Computer Science, Information Systems, (12/148). JCR-2017 F.I.: 4.305**.

5. Título: **Analysis of the evolution of the Spanish labour market through unsupervised learning**. Autores: **Fernando Nuñez-Hernández, José María Luna-Romera, María Martínez-Ballesteros, José C. Riquelme, Carlos Usabiaga**.

   En Revisión: **IEEE Access**, IEEE, ISSN: 2169-3536, Fecha de Publicación: en proceso, **Q1 en Computer Science, Information Systems, (23/155). JCR-2018 F.I.: 4.098**.

6. Título: **Big-Data Analysis for Demand Response in a Smart Electricity Market**. Autores: **José Antonio Fábregas, José María Luna-Romera, David Gutiérrez-Avilés, José C. Riquelme**.

   En revisión: **IEEE Access**, IEEE, ISSN: 2169-3536, Fecha de Publicación: en revisión, **Q1 en Computer Science, Information Systems, (23/155). JCR-2018 F.I.: 4.098**.

A.2. Otras Revistas

7. Título: **An approach to validity indices for clustering techniques in Big Data**. Autores: **José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros, José C. Riquelme**.

   Publicado en: **Progress in Artificial Intelligence**, Springer, ISSN: 2192-6352, Fecha de Publicación: Junio 2018, Volumen: 7, Issue 2, En Páginas: 91-94, DOI: https://doi.org/10.1016/10.1007/s13748-017-0135-3. **Q2 en Artificial Intelligence. SJR-2018: 0.513**. Citas Scopus-Scholar-WOS: 3-9-1.

A.3. Conferencias Nacionales

8. Título: **An Approach to Silhouette and Dunn Clustering Indices Applied to Big Data in Spark**. Autores: **José María Luna-Romera, María Martínez-Ballesteros, Jorge García-Gutiérrez,**

José C. Riquelme. Publicado en: **XVII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA). LNCS 9868**, ISBN: 978-3-319-44635-6, Fecha de Publicación: 2016, En Páginas: 160-169. DOI: 10.1007/978-3-319-44636-3_15. Citas Scopus-Scholar-WOS: 4-8-5.

9. Título: **Análisis Big Data para la Respuesta a la Demanda en el Mercado Eléctrico**. Autores: **Jose Antonio Fábregas, José María Luna-Romera, Ángel Arcos-Vargas, José C. Riquelme, Javier Tejedor**. Publicado en: **XVIII Conference of the Spanish Association for Artificial Intelligence (CAEPIA).**, Fecha de Publicación: Octubre 2018, En Páginas: 777-783.

10. Título: **Aproximación al índice externo de validación de clustering basado en chi cuadrado**. Autores: **José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros, José C. Riquelme**. Publicado en: **XVIII Conference of the Spanish Association for Artificial Intelligence (CAEPIA).**, Fecha de Publicación: Octubre 2018, En Páginas: 821-826.

11. Título: **Indexes to Find the Optimal Number of Clusters in a Hierarchical Clustering**. Autores: **José David Martín-Fernández, José María Luna-Romera, Beatriz Pontes, José C. Riquelme**. Publicado en: **14th International Conference on Soft Computing Models in Industrial and Environmental Applications. SOCO 2019. Advances in Intelligent Systems and Computing vol 950**, ISBN: 978-3-030-20055-8, Fecha de Publicación: Mayo 2019, En Páginas: 3-13. DOI: 10.1007/978-3-030-20055-8_1

A.4. Proyectos I+D+i

Esta tesis doctoral ha sido desarrollada dentro del contexto de los siguientes proyectos de investigación:

- Título: **Modelos Avanzados para el Análisis Inteligente de Información. Aplicación a Datos Biomédicos y Medioambientales**. Investigadores principales: **Cristina Rubio Escudero**. Entidad: **Junta de Andalucía. Consejería de Innovación, Ciencia y Empresas**. Periodo: **2013-2017**. Referencia: **P11-TIC-7528**.

- Título: **Big Time-Aware Data: Análisis de Datos Masivos Indexados en el Tiempo**. Investigadores principales: **José Cristobal Riquelme Santos**. Entidad: **Gobierno de España. Ministerio de Economía y Competitividad**. Periodo: **2015-2017**. Referencia: **TIN2014-55894-C2-1-R**.

- Título: **Big Data Streaming: Análisis de Datos Masivos Continuos. Modelos Descriptivos**. Investigadores principales: **José Cristobal Riquelme Santos. Cristina Rubio Escudero**. Entidad: **Gobierno de España. Ministerio de Economía y Competitividad**. Periodo: **2018-2020**. Referencia: **TIN2017-88209-C2-2-R**.

A.5. Estancias

A continuación se listan las estancias de investigación realizadas a lo largo de la elaboración de la tesis doctoral:

- Universidad: **Universidad de Granada**. Centro: Escuela Técnica Superior de Ingeniería Informática y de Telecomunicación. Fechas: **enero 2017 a junio 2017**
- Universidad: **Arizona State University** (Estados Unidos). Centro: Eyring Materials Center. Fechas: **febrero 2019 a mayo de 2019**

Bibliografía

[1] P.A. Alaba, S.I. Popoola, L. Olatomiwa, M.B. Akanle, O.S. Ohunakin, E. Adetiba, O.D. Alex, A.A.A. Atayero, and W.M.A. Wan Daud. Towards a more efficient and cost-sensitive extreme learning machine: A state-of-the-art review of recent trend. *Neurocomputing*, 350:70–90, 2019.

[2] Ramiz M. Aliguliyev. Performance evaluation of density-based clustering methods. *Information Sciences*, 179(20):3583 – 3602, 2009.

[3] A. K. Alok, S. Saha, and A. Ekbal. A min-max distance based external cluster validity index: Mmi. In *2012 12th International Conference on Hybrid Intelligent Systems (HIS)*, pages 354–359, 2012.

[4] M. Ankerst, M.M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 28(2):49–60, 1999.

[5] J. Arias, J.A. Gamez, and J.M. Puerta. Learning distributed discrete bayesian network classifiers under mapreduce with apache spark. *Knowledge-Based Systems*, 117:16–26, 2017.

[6] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382, 2005.

[7] Asa Ben-Hur and Isabelle Guyon. Detecting stable clusters using principal component analysis. In Michael J. Brownstein and Arkady B. Khodursky, editors, *Functional Genomics: Methods and Protocols*, pages 159–182. Humana Press, Totowa, NJ, 2003.

[8] V. Berikov and I. Pestunov. Ensemble clustering based on weighted coassociation matrices: Error bound and convergence properties. *Pattern Recognition*, 63:427–436, 2017.

[9] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27, 1974.

[10] D.N. Campo, G. Stegmayer, and D.H. Milone. A new index for clustering validation with overlapped clusters. *Expert Systems with Applications*, 64:549 – 556, 2016.

[11] M. Castro-Franco, M.A. Córdoba, M.G. Balzarini, and J.L. Costa. A pedometric technique to delimitate soil-specific zones at field scale. *Geoderma*, 322:101–111, 2018.

[12] D.L. Davies and D.W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2), 1979.

[13] R. Davoodi and M.H. Moradi. Mortality prediction in intensive care units (icus) using a deep rule-based fuzzy classifier. *Journal of Biomedical Informatics*, 79:48–59, 2018.

[14] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[15] Richard Dubes and Anil K. Jain. Clustering techniques: The user's dilemma. *Pattern Recognition*, 8(4):247–260, 1976.

[16] J.C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1), 1974.

[17] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, and Abdelaziz Bouras. A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3):267–279, 2014.

[18] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

[19] Pasi Fränti, Mohammad Rezaei, and Qinpei Zhao. Centroid index: Cluster level similarity measure. *Pattern Recognition*, 47(9):3034 – 3045, 2014.

[20] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. volume 37, pages 29–43, 12 2003.

[21] U. Ghia, K.N. Ghia, and C.T. Shin. High-re solutions for incompressible flow using the navier-stokes equations and a multigrid method. *Journal of Computational Physics*, 48(3):387–411, 1982.

[22] Leo A. Goodman and William H. Kruskal. *Measures of Association for Cross Classifications*, pages 2–34. Springer New York, New York, NY, 1979.

[23] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, 2001.

[24] Jiawei Han, Micheline Kamber, Jian Pei, Jiawei Han, Micheline Kamber, and Jian Pei. 10 – Cluster Analysis: Basic Concepts and Methods. In *Data Mining*, pages 443–495. Morgan Kaufmann, 2012.

[25] Julia Handl, Joshua Knowles, and Douglas B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

[26] C. Hennig and T.F. Liao. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 62(3), 2013.

[27] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[28] Sohail Jabbar, Abid Ali Minhas, Anand Paul, and Seungmin Rho. Multilayer cluster designing algorithm for lifetime improvement of wireless sensor networks. *The Journal of Supercomputing*, 70(1):104–132, 2014.

[29] Julien Jacques and Cristian Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

[30] Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[31] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 16–22, New York, NY, USA, 1999. ACM.

[32] Yang Lei, James C. Bezdek, Simone Romano, Nguyen Xuan Vinh, Jeffrey Chan, and James Bailey. Ground truth bias in external cluster validity indices. *Pattern Recognition*, 65:58 – 70, 2017.

[33] Chuan Liu, Wenyong Wang, Martin Konan, Siyang Wang, Lisheng Huang, Yong Tang, and Xiang Zhang. A new validity index of feature subset for evaluating the dimensionality reduction algorithms. *Knowledge-Based Systems*, 121:83 – 98, 2017.

[34] Ezequiel López-Rubio, Esteban J. Palomo, and Francisco Ortega-Zamorano. Unsupervised learning by cluster quality optimization. *Information Sciences*, 436-437:31 – 55, 2018.

[35] José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros, and José C. Riquelme Santos. An approach to validity indices for clustering techniques in big data. *Progress in Artificial Intelligence*, 7(2):81–94, 2018.

[36] José María Luna-Romera, María Martínez-Ballesteros, Jorge García-Gutiérrez, and José C. Riquelme. External clustering validity index based on chi-squared statistical test. *Information Sciences*, 487:1 – 17, 2019.

[37] J.D. Martín-Fernández, J.M. Luna-Romera, B. Pontes, and J.C. Riquelme-Santos. Indexes to find the optimal number of clusters in a hierarchical clustering. *Advances in Intelligent Systems and Computing*, 950:3–13, 2020.

[38] S. Mazumdar, D. Seybold, K. Kritikos, and Y. Verginadis. A survey on data storage and placement methodologies for cloud-big data ecosystem. *Journal of Big Data*, 6(1), 2019.

[39] Marina Meilă. Comparing clusterings by the variation of information. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines*, pages 173–187, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[40] Marina Meilă and David Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42(1):9–29, 2001.

[41] Zachary Miller, Brian Dickinson, William Deitrick, Wei Hu, and Alex Hai Wang. Twitter spammer detection using data stream clustering. *Information Sciences*, 260:64–73, 2014.

[42] F. Padillo, J.M. Luna, and S. Ventura. Exhaustive search algorithms to mine subgroups on big data using apache spark. *Progress in Artificial Intelligence*, 6(2):145–158, 2017.

[43] A. Paul, A. Ahmad, M. M. Rathore, and S. Jabbar. Smartbuddy: defining human behaviors using big data analytics in social internet of things. *IEEE Wireless Communications*, 23(5):68–74, 2016.

[44] Rubén Pérez-Chacón, José M. Luna-Romera, Alicia Troncoso, Francisco Martínez-Álvarez, and José C. Riquelme. Big data analytics for discovering electricity consumption patterns in smart cities. *Energies*, 11(3), 2018.

[45] S. Ramirez-Gallego, H. Mourino-Talin, D. Martinez-Rego, V. Bolon-Canedo, J.M. Benitez, A. Alonso-Betanzos, and F. Herrera. An information theory-based feature selection framework for big data under

apache spark. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(9):1441–1453, 2018.

[46] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[47] M. Rezaei and P. Fränti. Set matching measures for external cluster validity. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2173–2186, 2016.

[48] J. Rodríguez, M.A. Medina-Pérez, A.E. Gutierrez-Rodríguez, R. Monroy, and H. Terashima-Marín. Cluster validation using an ensemble of supervised classifiers. *Knowledge-Based Systems*, 145:1–14, 2018.

[49] J.C. Rojas-Thomas, M. Santos, and M. Mora. New internal index for clustering validation based on graphs. *Expert Systems with Applications*, 86:334 – 349, 2017.

[50] P.J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(C), 1987.

[51] Tomer Sagi, Avigdor Gal, Omer Barkol, Ruth Bergman, and Alexander Avram. Multi-source uncertain entity resolution: Transforming holocaust victim reports into people. *Information Systems*, 65:124–136, 2017.

[52] S. Saleti and S. R.B.V. A mapreduce solution for incremental mining of sequential patterns from big data. *Expert Systems with Applications*, 133:109–125, 2019.

[53] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194, 1998.

[54] Beatriz Sevilla-Villanueva, Karina Gibert, and Miquel Sànchez-Marrè. Using cvi for understanding class topology in unsupervised scenarios. In Oscar Luaces, José A. Gámez, Edurne Barrenechea, Alicia Troncoso, Mikel Galar, Héctor Quintián, and Emilio Corchado, editors, *Advances in Artificial Intelligence*, pages 135–149, Cham, 2016. Springer International Publishing.

[55] R.R. Sokal and P.H.A. Sneath. *Principles of Numerical Taxonomy*. Books in biology. W. H. Freeman, 1963.

[56] Apache Spark. Apache Spark, Lightning-fast cluster computing. `https://spark.apache.org/`, 2017.

[57] Apache Spark. Clustering - Spark 2.2.0 Documentation. `https://spark.apache.org/docs/2.2.0/ml-clustering.html`, 2018.

[58] R. Talavera-Llames, R. Pérez-Chacón, A. Troncoso, and F. Martínez-Álvarez. Mv-kwnn: A novel multivariate and multi-output weighted nearest neighbours algorithm for big data time series forecasting. *Neurocomputing*, 353:56–73, 2019.

[59] Robert Tibshirani and Guenther Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.

[60] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, 2000.

[61] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1073–1080, New York, NY, USA, 2009. ACM.

[62] Wei Wang, Jiong Yang, and Richard R. Muntz. Sting: A statistical information grid approach to spatial data mining. In *Proceedings of the 23rd International Conference on Very Large Data Bases*, VLDB '97, pages 186–195, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[63] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and E Y Chang. Parallel Spectral Clustering in Distributed Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):568–586, 2011.

[64] Junjie Wu, Hui Xiong, and Jian Chen. Adapting the right measures for k-means clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 877–886, New York, NY, USA, 2009. ACM.

[65] Hamdi Yahyaoui and Hala S. Own. Unsupervised clustering of service performance behaviors. *Information Sciences*, 422:558 – 571, 2018.

[66] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauly, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 15–28, San Jose, CA, 2012. USENIX.

[67] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: An efficient data clustering method for very large databases. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 25(2):103–114, 1996.

[68] Yaqian Zhang, Jacek Mańdziuk, Chai Hiok Quek, and Boon Wooi Goh. Curvature-based method for determining the number of clusters. *Information Sciences*, 415-416:414 – 428, 2017.

[69] B. Zhao and J. Wang. Unification of particle velocity distribution functions in gas-solid flow. *Chemical Engineering Science*, 177:333–339, 2018.

[70] Ying Zhao and George Karypis. Criterion functions for document clustering: Experiments and analysis. Technical report, University of Minnesota, Department of Computer Science, Minneapolis, 2001.